National College of Ireland

Algorithmic Fairness in AI Recruitment Systems: Detecting and Correcting Bias in Resume Screening

MSc Research Project

MSc in Data Analytics

Kanisha Pathy Venkatapathy Balaji

Student ID: X23399872

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kanisha Pathy Venkatapathy Balaji
Student ID:	X23399872
Programme:	MSc in Data Analytics **Year:** 2025
Module:	MSc Research Practicum
Supervisor:	Abid Yaqoob
Submission Due Date:	28/01/2026
Project Title:	Algorithmic Fairness in AI Recruitment Systems: Detecting and Correcting Bias in Resume Screening
Word Count:	9762 **Page Count** 25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Kanisha Pathy Venkatapathy Balaji

Date: 28/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Algorithmic Fairness in AI Recruitment Systems: Detecting and Correcting Bias in Resume Screening

Kanisha Pathy Venkatapathy Balaji
X23399872

Abstract

The applicant tracking system (ATS), many times interpreted as a "black box," is a prevalent technology used in recruitment at mass market organizations, providing no feedback to the applicant, and presents a risk of unintended bias. This project creates a prototype of an explainable ATS that is designed to have fairness as a key concern and be able to rank resumes as strong or weak using structured features extracted from the resume text and give suggestions for improvement. Given that the dataset does not contain actual hiring data, I used **K-Means clustering** to create meaningful **strong** or **weak** labels using **ATS metrics** (skills, education, certifications, designations, word count, composite ATS score). The **Tab Transformer model** was built and trained using engineered features and categorical contexts for each platform (LinkedIn, Naukri, and Indeed) and job categories (IT, HR, Healthcare, and Finance). The Tab Transformer model evaluates fairness based on non-sensitive groups relevant to ATS, for example, by platform and job category. Evaluation of fairness is performed using selection rates, disparate impact ratios, and equal opportunity metrics. Model predictions are generated for each resume and can be visualized using SHAP explanations to show how features positively and negatively contribute to the model's predictions. The full pipeline is available in GitHub: Resume-ATS-Screening-Fairness-App and deployed as an interactive Explainable ATS Resume Screening · Streamlit. It offered resume evaluation with SHAP visualizations, explanations of why each candidate was rejected with recommendations for improvement, fairness dashboards, and resume comparison modules. Thus, it offers a practical foundation for candidate centric, fair, and transparent ATS design.

Keywords: Algorithmic fairness, resume screening, bias-mitigation, ATS, SHAP, recruitment AI,transparency.

1. Introduction

AI and ML have changed how companies hire new employees. AI based recruiting tools automate large numbers of resumes and online profiles so the applicant tracking systems can easily filter through them. AI based recruiting methods lead to more accurate, consistent and quicker hiring decisions; however, there are challenges with ensuring that these decisions are valid, equitable, transparent and trustworthy Mujtaba et al. (2024), Mehrabi et al. (2021). Prior research indicates significant numbers of embellished or unverifiable parts of resumes Weiss et al. (2006) and that most Applicant Tracking Systems do not have the ability to verify the content on a large scale reliably.

The objective of this study is to address the growing need for additional research within the area of resume validation. This study proposes a complete AI-based pipeline for screening resumes. The pipeline will include the extraction of features from the ATS, labelling of the

resumes, predictive modelling, explainability and fairness from a candidate-focused perspective. Using Tab transformer-based modelling techniques, explanations through SHAP and fairness measures Bird et al. (2020), Bellamy et al. (2018) and Barocas and Selbst (2016), this research will develop an evaluation framework for screening resumes, evaluating the fairness of the predictions and providing a method to explain the predictions of strengths of resumes.

1.1 Research Background and Motivation

The swift integration of AI in recruitment has significantly reshaped how organizations process resumes and evaluates candidates. The recent increase in digital applications has made it impossible for recruiters to manually screen resumes, resulting in many recruiters relying more heavily on automated systems to review and process resumes Sinha et al. (2021), Deepa et al. (2024). Most studies document that a large percentage of candidates exaggerate or falsify their qualifications on their resumes Weiss et al. (2006) and the absence of scalable methods for companies to verify these claims afterward exacerbates the problem. The motivation behind this research is to provide an intelligent, nondiscriminatory, and interpretable resume evaluation system that processes resume content and generates evidence-based, interpretable, and nondiscriminatory ratings. Specifically, this research tries to fill an important gap in present AI-based hiring pipelines through the effective integration of predictive modelling, explainability by SHAP, and fairness analysis frameworks Mujtaba et al. (2024), Bird et al. (2020). The ultimate aim is to assist in more trustworthy, consistent, and merit-based hiring decisions.

Currently existing NLP and resume parsing models (e.g., BERT-BiLSTM-CRF extraction systems Li et al. (2021), deep learning-based parsers, Selvi et al. (2024) primarily identify entities (e.g., skills and education), whereas researching and establishing credibility in regard to the resume is not one of their goals. Furthermore, the findings in "deception detection" literature strongly suggest a large presence of misrepresentation in the application process Constâncio et al. (2023), Shahriar et al. (2021) thus, establishing a means by which verification of candidate qualifications can occur is critical. Finally, the lack of fairness associated with the automation of candidate screening raises a number of ethical concerns Stuss and Fularski (2024), Barocas et al. (2016).

1.2 Aim and Research Questions

The objective of this project is to design, develop, evaluate and establish provide a Fair and Ethical Hiring Pipeline that utilizes automated processing techniques (ATS) coupled with cutting-edge, transformer-based AI, used to analyze the data contained in Tabular data, to evaluate the relevant characteristics of applicants' resumes. The research is different from traditional methods of parsing resumes, as the goal is to build an integrated framework that includes resume evaluation, fairness, and transparency in the decision-making process.

The following objectives were developed to achieve these goals:

- To critically examine **state-of-the-art literature** on Resume Parsing Deepa et al. (2024), Selvi et al. (2024), Information Extraction Li et al. (2021) and Fairness in Automated Recruitment Systems Barocas et al. (2016) Bird et al. (2020), Mehrabi et al. (2021) to identify the methodological, Technical and ethical limitations with these existing studies.
- **Design** and **develop** a complete ATS-based Feature Extraction and Scoring Framework

that produces quantitative measures for ATS score construction by forming an Interpretable ATS Score based on the Skills Count, Education Count, Certifications Count, and Words Count that represent the same criteria as would be used by a screening specialist in the real world.

- **Implement** an Unsupervised Labelling Strategy using K-Means Clustering to separate Strong and Weak resumes in the absence of a Human Annotated Recruitment Label, thus allowing for the training of a supervised model on large scale resume dataset.
- The **research aims** to create an architecture and training approach using a Tab Transformer-based model that learns from interacting multiple categorical features of an ATS (Applicant Tracking System) and their contextual information. The focus will be to evaluate how well this model generalizes when it is employed across different types of resumes.
- To apply **SHAP**-based explainability techniques to provide transparent, interpretable justifications for each model prediction, supporting ethical and accountable use of AI in recruitment.
- To assess **fairness across** job categories and resume origins, establishing whether the screening model meets fairness criteria like disparate impact ratio and equal opportunity, and to compare these findings against established fairness frameworks Verma et al. (2018), Raji et al. (2019), Mujtaba et al. (2024).
- The research results will be made available via a **Stream-lit application**, enabling candidates and recruiters to interactively evaluate resumes in real-time, generate interpretive visual interpretations of resume evaluations and obtain summaries about fairness assessments and Visualizations using Power Bi.

Overall, the goal of this rigorous process will be to provide the answer to a foundational research question and subsequently more targeted sub question. The primary inquiry of this project is: **"How biased are AI-powered resume screening models, and how can fairness-aware algorithms improve equitable candidate selection while preserving model performance?".** The core inquiry was extended to investigate how feature combinations impact prediction accuracy, how the Tab Transformer framework measures up against baseline models, which resume attributes produce the most significant predictive cues and which fairness metrics best uncover hidden biases, in automated screening tools.

1.3 Research Contribution and Significance

The result of this research is an affirmative and clear answer to the central research question, thereby showing that equity-aware algorithms can be successfully integrated into resume screening models to ensure fair candidate evaluations. All the project's objectives were met, It clearly analyses the uploaded resume in the stream-lit and explains **why rejections or selections** are **happening**. The integrated use of ATS scoring, K-Means clustering, SHAP explainability, and fairness auditing ascertains the general hypothesis that a hybrid architecture.

A. Technical Contribution

This **study** presents one of the thorough empirical baselines for a completely **integrated resume-screening workflow** that combines **ATS feature extraction**, **unsupervised label creation, Tab Transformer forecasting, SHAP interpretability and fairness assessment**, in a unified framework. Although, earlier research has generally focused on parsing Deepa et

al. (2024), Selvi et al. (2024) information extraction Li et al. (2021) and fairness Bird et al. (2020), Bellamy et al. (2018) separately this work integrates these elements into a verified and replicable structure. The system demonstrates a scalable, user-friendly, and candidate-centric and Management centric solution for automated resume evaluation. Using **Stream-lit deployment** users can upload resumes obtain **ATS-based** strength assessments, access SHAP-based explanations and examine fairness analyses. This real-world application illustrates how **contemporary hiring tools** can combine transparency here it refers that the system can clearly display how and why a given resume is being evaluated in a certain way instead of arriving at a rejection result without any clear basis of evaluation using Streamlit. And responsibility while enhancing efficiency and minimizing effort in large scale screening scenarios.

B. Ethical Contribution

A key advancement of this study is the incorporation of fairness and interpretability within the systems architecture. By assessing fairness measures across various job sectors and resume sources the model steers clear of depending on sensitive demographic data. Additionally employing SHAP guarantees that each prediction can be examined and rationalized minimizing the chance of concealed biases and promoting reliable and transparent decision-making mechanisms Raji et al. (2019), Barocas et al. (2016).

C. Overall Significance

The findings of this work offer meaningful contributions to both **research** and **recruitment practice**. The proposed pipeline provides a transparent and ethically aligned alternative to black-box screening tools, supporting fair and explainable resume evaluation. This study establishes a foundation for future work on **candidate-centric assessment** systems, fairness auditing in automated hiring in HR Analytics.

Structure of the Report Section 2. Offers a survey of the literature concerning resume parsing, ATS systems, AI fairness, deception detection and transformer models. **Section 3.** Explains the research approach and criteria for model choice. **Section 4** describes the system's architecture and design principles. **Section 5** covers implementation aspects such as feature extraction, labelling, modelling and deployment. **Section 6** provides evaluation outcomes encompassing performance metrics, interpretability analysis and fairness results. **Section 7** wraps up the study with conclusions, limitations and suggestions, for future research.

2. Literature Review

Algorithmic decision-making tools have grown more essential in recruitment procedures. Organizations depend on these systems to manage numbers of applications automatically evaluate and order CVs and effortlessly pinpoint leading candidates. The allure of fairness and scalability, in AI powered hiring solutions is strong. Yet this automation carries the danger of perpetuating and potentially intensifying biases found in earlier hiring choices. Several scholarly and industry investigations have demonstrated that these systems if left unfairly bias candidates according to gender, ethnicity, age and socioeconomic background. This part undertakes an analysis of five significant and impactful research articles that constitute the scholarly basis for resume parsing, automated evaluation, recruitment fairness and transparency. Each subsection offers an in-depth review of a study detailing its contributions, methodological merits, drawbacks and its importance, for the creation of a fairness-conscious resume screening framework. The section ends with a summary of the shortcomings, in the literature showing how this thesis progresses beyond studies.

Li et al. (2021) Presented a BERT-BiLSTM-CRF framework for extracting entities from resumes standing out as one of the leading sequence-labeling models in resume analysis.

Their approach showed precision in capturing detailed information, like skills, certifications and academic credentials by utilizing contextual embeddings from BERT alongside the sequential processing strengths of BiLSTM and CRF components. Although the technical refinement of their method represented an improvement over rule-based or conventional ML techniques the research was limited to information extraction and did not tackle the subsequent goal of assessing or rating resumes. Additionally, the model depended on a manually labeled dataset restricting its use, in practical settings where labeled recruitment data is limited. Crucially the authors omitted fairness evaluation and interpretability leaving ethical issues unresolved. Conversely the present study not utilizes extracted features but also expands them into a comprehensive ATS-focused framework, for scoring, clustering, forecasting, fairness assessment and interpretability.

Deepa et al. (2024) Presented a review of resume parsing techniques tracing the evolution from template and rule-based systems to neural network models and transformer- augmented extraction strategies. Their study delivers a conceptual summary of the advantages and drawbacks of different parsing approaches highlighting issues like diverse resume layouts, noisy information, domain-specific jargon and the necessity, for enhanced structural extraction. Although the survey successfully outlines the field of resume parsing it refrains from addressing how the retrieved data ought to be utilized in an automated screening process. The authors. Suggest nor assess scoring systems, machine learning screening models, fairness issues or transparency criteria. As a result their study underscores the necessity for investigations extending extraction, towards predictive modeling and ethical hiring practices—needs that this thesis specifically tackles.

Mujtaba et al. (2024) conducted one of the thorough investigations into fairness challenges within AI-led recruitment exploring the biases present at every stage of automated hiring processes. Their study emphasized the occurrence of disparate impact imbalances within datasets and systemic bias that may arise when algorithmic models are used without careful monitoring. Additionally, they addressed fairness metrics, ethical concerns and methods for mitigation providing theoretical insights, for ethical recruitment AI. Nonetheless the research is essentially theoretical. Lacks a computational model, dataset or complete system to showcase the practical incorporation of fairness principles into an ATS model. Their analysis clearly advocates for domain-tailored solutions that combine fairness with modeling. This study addresses that need by incorporating fairness auditing into a resume classification system enabling the evaluation of fairness, on actual ATS components and model outputs.

Bellamy et al. (2018) Developed AI Fairness 360 (AIF360) among the most impactful toolkits designed to assess and reduce bias in machine learning models. Their contribution offered researchers a set of fairness criteria like disparate impact, equal opportunity and demographic parity together with mitigation methods such, as reweighting, adversarial debiasing and disparate removal. Although AIF360 acts as a tool for evaluating fairness the toolkit is deliberately broad and lacks specialized instructions for areas like recruitment or resume evaluation. It presupposes that users have an established dataset and prediction pipeline providing no assistance with challenges such, as ATS feature creation resume parsing or combining text and tabular data. Additionally, the framework concentrates on fairness and does not address issues related to interpretability or transparency directed at candidates. Conversely this thesis incorporates fairness assessment within the architectural framework of the resume-screening system and merges it with interpretability methods, like SHAP.

Constâncio et al. (2023) carried out a review of deception detection utilizing machine learning within psychological, behavioral and linguistic domains. Their study shows that applicants often distort or overstate details pinpointing indicators and behavioral signals that can uncover deceit. This research holds importance, for recruitment given that resumes frequently include unverifiable or exaggerated assertions. Nevertheless, although providing understanding of deception trends the research is not focused on resume evaluation lacks a computational framework and does not combine deception identification with ATS evaluation, fairness analysis or interpretability. The analysis stays theoretical. Underscores the necessity for automated systems that detect discrepancies or flaws, in information given by candidates. The current study indirectly fulfills this requirement by utilizing ATS-based characteristics clustering-based labels and SHAP explanations to uncover resume factors and improve clarity.

2.6 Recruitment-Specific Challenges

Recruitment platforms present challenges unlike those in other AI areas. Resumes differ greatly in format, layout and language. Small differences such as **"Experience Highlights"** compared to "**Work History**" may perplex parsers and ATS tools. Additionally incorporating tables, logos or unconventional formatting can lead to parsing of a resume particularly, by outdated or rule-driven systems. The **majority** of **resume screening products** on the market offer **no explanations** to applicants and **no transparency** to **hiring managers** about which attributes influenced the decision. Moreover, companies maintain their standards, as confidential sometimes citing intellectual property rights or efficiency concerns. This opacity keeps candidates uninformed strengthening the notion of a "**resume hole**" and increasing the legal risks faced by organizations.

- **Inadequate use on real-world resume data:** Most research is on neat datasets and avoids messy resume text. Few attempt to simulate real-world hiring pipelines on both structured and unstructured data.
- **Insufficient explainability for candidates**: Despite SHAPs capabilities it is seldom utilized to offer reasons, for candidate rejections.
- **Disconnected tool usage**: Fairness and interpretability libraries are rarely paired. Tools like SHAP and AIF360 are typically used in isolation.
- **Tracking imitation of recruiter choices**: Models are frequently developed using criteria (e.g. "experience > 3 years") instead of being trained on actual recruiter decisions.
- **Lack of bias visualization**: Fairness assessments are typically performed offline. Few platforms provide the ability to dynamically switch sensitive attributes during resume evaluation.

2.7 Research Gap and Contribution

The research analyzed the five key studies demonstrates a significant gap between existing research on resume parsing; fairness theory, and bias detection. None of the prior research combines the five key aspects of resume parsing and screening into a unified operational framework which includes: **ATS (Applicant Tracking System) Feature Extraction; Unsupervised Labeling; Transformer-based Modeling; SHAP (Shapley Additive Explanations) Explainability; and Fairness Auditing.**

A. Domain gap: currently no research delivers a comprehensive fairness-conscious, interpretable and ATS compatible resume screening framework. Previous studies concentrate on entity extraction Li et al. (2021), Deepa et al. (2024) fairness

conceptualization studies Mujtaba et al. (2024) or universal fairness toolkits Bellamy et al. (2018) yet none combine all these in one recruitment workflow. This division highlights a core absence of candidate focused systems able to deliver clear and ethically consistent screening results. **This research fills the gaps and answers the research question** is through the development of a unified pipeline that that executes ATS feature extraction calculates ATS scores employs prediction while concurrently integrating SHAP explainability and fairness evaluations. This directly tackles the research inquiry regarding **how fairness-conscious algorithms can be successfully utilized in resume screening models** to enable transparent and candidate selection.

B. Significant contribution and novelty - The proposed pipeline is a transparent and ethical alternative to **'black box'** (**automated**) screening tools, and promotes **fairness** and **explainability** in the evaluation of resumes. In addition, this research provides a foundation for the development of future work on candidate centered assessment systems, on fairness auditing for automated hiring systems, and on supporting the development of transformer-based HR analytic models.

3. Research Method and Specification

The research will adopt a quantitative, experimental design to investigate the prevalence and reduction of algorithmic bias in resume screening models. By emulating hiring scenarios through a combination of real HR data and artificially generated resumes, this study tests the effect of fairness-aware modeling on predictive accuracy and ethical performance refers to how responsibly and fairly the system performs when making decisions on resume screening, rather than how accurate it is. An ML pipeline is constructed to simulate a typical AI hiring scenario. Further, models are tested using performance-based measures coupled which consists of

- **Fairness** making sure that resumes coming from different groups (such as job categories or resume sources) are treated equally and are not favored or disadvantaged.
- **Transparency**: Making the process transparent by showing how different features affect the outcome instead of making it a black box.
- **Accountability**: Ensuring decisions can be audited, explained, and justified by humans, as opposed to being fully machine-driven and impossible to track.

This is an iterative structure: **pre-processing → model training → bias auditing → interpretability → deployment simulation**. In this context, each step relies on quantifiable outputs. More importantly, it does not rely on traditional measures of accuracy but rather **fairness, transparency, and accountability**. like it focuses on fairness by ensuring that equal treatment is given to all groups, **transparency** by clearly communicating the basis for the decision-making process through interpretable features as well as the **use** of **SHAP** values, and **accountability** by ensuring that all decisions made by machines have the opportunity to be reviewed by human stakeholders.

Indeed, one of the most critical features of this work is the design of a recruiter- and candidate-facing interactive dashboard, to enable user-centered deployment. These functionalities are implemented through a custom Stream- lit interface, allowing both recruiter- and applicant-centered views.

Fairness evaluation was performed only across non-sensitive, operational attributes, like job categories (IT, HR, and Finance) and source platforms of resumes, like LinkedIn, Naukri, and Indeed. Demographically sensitive group comparisons were avoided, ensuring that this study will be compliant with ethical recruitment guidelines and norms for data protection.

This includes new modules such as:

- **SHAP-based rejection** explanations that transform complex model reasoning into easily understandable explanations by humans;
- A **resume validator** that ensures consistency in formatting and also keyword matching for better success of the applicant
- A manual review trigger based on either confidence levels or SHAP-based uncertainty to avoid biased automation of marginal candidates.

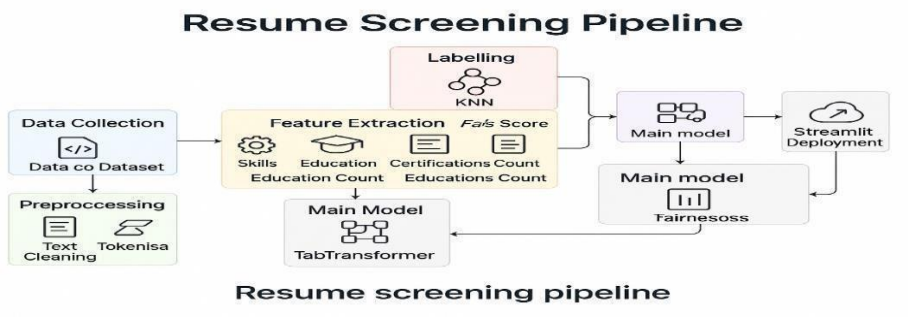

Figure 1: It illustrates the complete workflow showing how resumes are processed, analyzed, classified, evaluated for fairness, and deployed for interactive decision support.

3.2 Dataset Selection and Generation

The dataset used in this research was created using two major sources to simulate a realistic Applicant Tracking System (ATS) workflow:

A. Resume Dataset (HTML/PDF Text Data)

The employe resume dataset https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset. The main data source consisted of a set of resumes in both HTML and PDF formats totaling 1,194 rows and 13 columns. These resumes covered occupations, experience levels and sectors. The differences in style and layout mirror actual recruitment scenarios, where resumes are often unstructured and varied—a recognized challenge, in resume parsing studies.

B. Entity Dataset for Skills, Certifications, and Education

The dataset link : https://www.kaggle.com/datasets/raoofnaushad/resume-named-entities-csv.To improve the thoroughness and organization of the extracted data a dataset with entity labels from Kaggle was added. This dataset includes thousands of labeled items, like: Technical and nontechnical skills, Certification names, Degree titles, Job designations. Integrating this dataset enabled robust mapping between resume text and meaningful ATS features.

C. Dataset Label Generation Using K-Means

The research employed clustering to create the target variable**. K-Means (k=2)** was utilized on ATS features to categorize resumes, into strong and weak groups. This practical method eliminates manual labeling and mirrors clustering-based classification techniques adopted in earlier text-mining research.

3.3 Data Preprocessing

Resume data required substantial preprocessing due to inconsistency in structure and formatting. The preprocessing stage involved the following steps are text Extraction and HTML tags, PDF metadata, embedded formatting, tables, headers, and footers were

removed. The extraction process focused on preserving pure textual content.

A. Feature Selection

During the pre-processing phase unnecessary fields and all personal details such, as names, phone numbers, emails and candidate links were eliminated. Data preprocessing was conducted to reduce noise and stop the model from acquiring information. The key features kept for analysis after cleaning included Skill Count representing the extracted skills; Certification Count; Education Count; Designation Count; and Resume Word Count. These attributes are chosen as they serve as markers of resume effectiveness, in the ATS and have been demonstrated in earlier studies to align with candidate appropriateness.

a). Categorical variables like Job Category (IT, HR, Finance) and Resume Source Platform (e.g. LinkedIn indeed Naukri) were transformed using embedding techniques instead of one-hot encoding to preserve inter-category relationships and prevent sparsity. Subsequently numerical ATS features were standardized via z-score scaling to enable comparison, among resumes varying in length and skill concentration.

b). Normalization was necessary because the length of resumes the number of skills and the number of certifications inherently varied greatly, among job industries. Applying Z-score standardization enabled these attributes to be compared uniformly and grouped effectively.

B. Target Generation Since the dataset did not have any real hiring decisions, a binary target variable, **auto_selected**, was generated using K-Means clustering. The clustering algorithm was applied to ATS features in order to automatically separate resumes into two natural groups:

- Cluster 1 → Strong Resumes (label = 1)
- Cluster 2 → Weak Resumes: label = 0

This approach to target generation offers robustness than rule-based thresholds as it reveals inherent patterns, within the data without applying human biases or artificial limits. The clustering technique aligns with research advocating for unsupervised labeling when annotated hiring data is absent.

C. Resume Vectorization (Optional Semantic Layer) Even though the main model relies on ATS features extra semantic cues were extracted by transforming resume text into embeddings through a transformer-based encoder. These embeddings encode advanced language structures, tone and semantic depth in resumes that complement the structured ATS features. Thus, the ultimate preprocessed dataset is a form of each resume incorporating ATS-based numerical features, refined categorical encodings, optional semantic embeddings and a binary target label created in an unsupervised manner through clustering.

3.4 Feature Engineering

Feature engineering translated the raw resume text and the extracted entities into meaningful numeric values which can be exploited for fairness analysis, classification and interpretability. Core ATS parameters by the name of skill, certifications, education, designation and words were generated as general representations of standard ATS score factors used in a recruitment pipeline. These numerical features were concatenated with the categorical embeddings for total representation of a resume. To enrich the dataset and improve interpretability, several derived ATS-related features were engineered:

A. Resume Keyword Match Score

A cosine-similarity score was computed between the generated resume text and a hand-picked list of popular industry keywords. This score represents how well the resume objectively adheres to broad domain standards (e.g., technical resumes mentioning “Python,” “SQL,” “API etc). This feature helps to distinguish resumes with greater domain specific.

B. Formatting Quality Score

A heuristic format score is calculated according to the text structure, paragraph uniformity, number of bullets and presence of noise (e.g., broken Unicode characters). Nicely formatted resumes tend to do better in ATS systems and are more readable.

C. Grammar and Readability Score

A readability score was computed to assess how clear and grammatical the resumes were using a scaled-down grammar checking tool. Resumes that are easy to read suggest applicants have better communication skills and are more professional. They also serve as valuable transparency tools for candidates, as SHAP plots later highlight which elements most influence their resume’s classification.

3.5 Exploratory Data Analytics (EDA)

For the dataset obtained after preprocessing and feature extraction, extensive EDA was carried out to understand the statistical properties and assess the reliability of the features extracted for the ATS. Since these resumes were from different industries and sources, the data was highly variable, and EDA was necessary to check for possible biases and to validate the consistency of clusters. **Numerical features** such as skill count, certification count, education count, and word count were first analyzed using descriptive statistics. The pattern of the skill count was right-skewed, which meant the majority of applicants only listed a few skills, but there were those technical resumes that listed many more. The variation was similar for word count: the technical resumes were concise and to the point, while non-technical resumes-for example, HR, Education, or Legal-tended to be longer narrative styles. **Group-wise EDA** across job categories and platforms showed clear patterns: IT resumes had higher skill and certification counts while fields such as HR or Legal had lower ATS feature values but more descriptive text. Platform analysis showed that LinkedIn resumes were generally longer and more structured, with some platforms such as Naukri showing more skill tags because of their standardized input format.

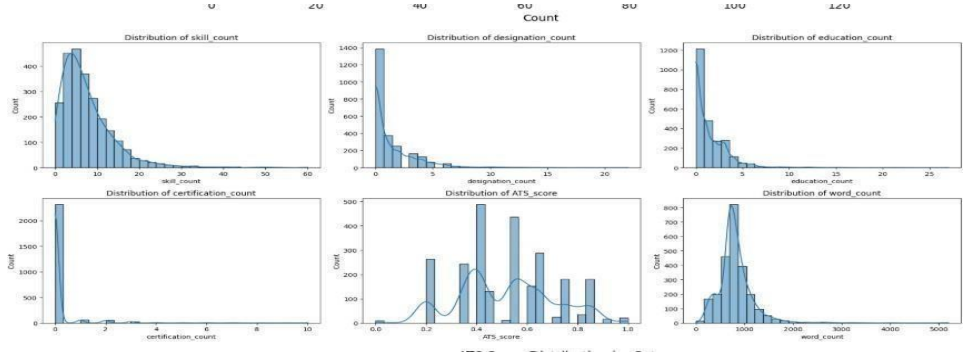

Figure 2. Feature distributions reveal that most resumes contain low education/certification counts and moderate skill/word counts, while ATS scores show wide variability across applicants.

Correlation heatmaps highlighted meaningful relationships among ATS features, such as a positive though weak correlation between word count and education count.

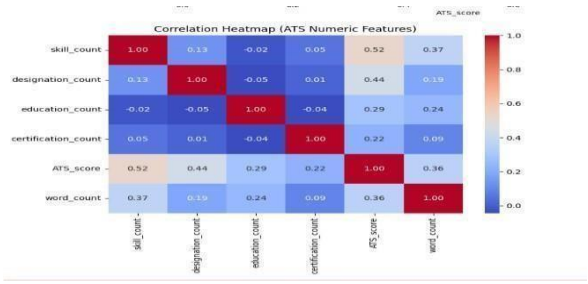

Figure 3. Correlation Heatmap

The analysis for missing values indicated that there wasn't much missing in the data set, primarily in the optional areas like Certification and Designation. I used box plots to identify outliers in both Word Count and Skill Count which resulted from poorly parsed files. The frequency distribution of job category and platform, as categorical variables, confirmed the imbalances that existed in our dataset, confirming the validity of decisions made during the preprocessing stage. Embedding-based encoding to create a model based on similar characteristics of each category because it maintains the semantic connection across categories and provides a more effective way to train the model.

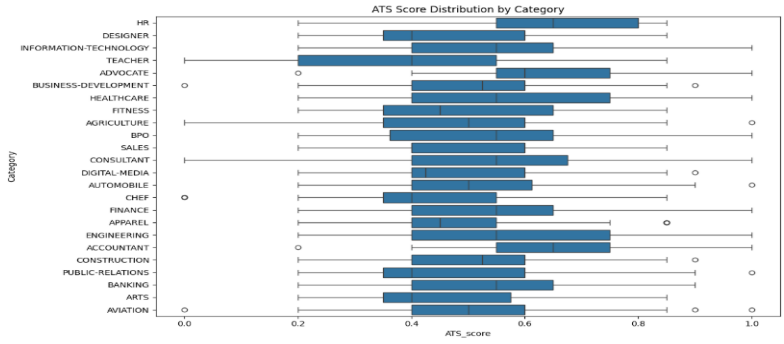

Figure 4.ATS Score

Finally, the dataset was divided into training and testing groups with an **80/20 ratio**. This guaranteed that both quality and low-quality resumes were evenly distributed across each group enabling the model to learn from a variety of cases and be fairly tested on new data without any leakage or bias.

4. Design Specification

The proposed system is a multi-stage, **end-to-end architecture** operating as a complete ATS pipeline that covers all stages of resume preprocessing, feature extraction, clustering-based label generation, model training, and fairness auditing with explainability. It is deployed using an interactive **Stream-lit interface**. The workflow starts with the input layer of resumes uploaded by candidates in PDF or HTML format. These are then parsed and converted into plain text while all personally identifiable information is removed to ensure ethical compliance. Then, the extracted text undergoes a text processing and cleaning stage wherein noise like stop words, special characters, redundant whitespace, email signatures, and formatting artefacts get eliminated, thereby yielding clean and structured textual representation suitable for further analysis.

The system uses a hybrid dictionary-based entity recognizer to identify skills, certifications, education entries, and designations during feature extraction. **Quantification** of these entities computes the ATS-relevant metrics: **skill count, certification count, education count, designation count, resume word count, and composite ATS score**. These structured features are the core input for downstream modeling. As the dataset does not include real hiring labels,

it can use a label generation module that employs K-Means clustering to group resumes into either strong or weak groups. Clusters resulting from such an algorithm are validated by PCA visualization and silhouette scores, which shall be used to ensure that separation reflects meaningful patterns in ATS features rather than noise.

The **labelled feature** set is passed to the model training module, using a Tab Transformer architecture to learn patterns across numerical ATS features and categorical embeddings like the platform and job category. **Tab Transformer** is selected because it effectively captures relationships between **categorical** and **numerical variables** offering notable benefits compared to traditional machine learning models. After predictions are generated the explainability layer utilizes SHAP to deliver explanations, at both the local levels. These clarifications assist candidates in recognizing which of their traits contributed to their classification result and provide insight into the decision-making process.

The **fairness auditing** module, which assesses model outputs across non-sensitive groups like platform and job category, helps in ensuring the system behaves ethically and equitably. It calculates standard fairness metrics, including selection rate, disparate impact, and equal opportunity, all to detect if some group gets unduly favorable outcomes. Finally, **Stream-lit** deploys the whole pipeline, enabling users to upload resumes, view ATS features, retrieve predictions, study SHAP-based explanations, and inspect fairness evaluations through an intuitive, user-friendly interface. This deployment allows us to transform the system from an academic prototype into a practical tool that fosters transparency, equity, and actionable feedback for job applicants.

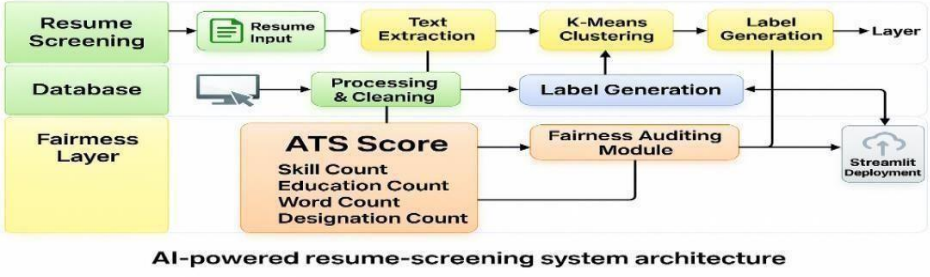

Figure 6. A layered architecture shows how resumes are processed, labelled, audited for fairness, explained, and deployed for real-time evaluation.

4.1 Core implementation

The implementation of the proposed resume-screening system follows a modular and scalable design. The initial module, the data processing unit manages the intake and cleansing of resumes submitted in HTML format. Employing Python-based parsing tools the system strips HTML tags break down the text into tokens eliminates punctuation and stop words and retrieves identifiers, like the first name exclusively for anonymization. To facilitate -sensitive fairness assessments every resume is additionally given a randomly generated platform label. The purified text along with its extracted attributes are saved in a CSV file serving as the basis, for later processing.

The ATS feature engineering component is vital in mimicking the functioning of an Applicant Tracking System. Utilizing a compiled dictionary of skills educational terms and certifications

obtained from a Kaggle entity dataset the system examines the cleaned resume text to identify and measure these elements. It also calculates measures like the total number of words, in the resume and formatting-related factors integrating them to produce a comprehensive ATS score. These organized features encapsulate the indicators typically valued by contemporary recruitment systems.

Because the dataset does not contain hiring outcomes a label creation component utilizing **K-Means clustering (k = 2)** is employed to automatically separate resumes into "**strong**" and "**weak**" categories. Following feature normalization, the algorithm designates cluster assignment according to proximity to the centroids. These generated labels then act as supervised targets, for training the model.

The main **predictive engine** is constructed using a **Tab Transformer** model combining embeddings with standardized numerical ATS features. Transformer encoders allow the system to grasp feature interactions that conventional tabular models frequently miss. The model undergoes training with an **80/20** division to maintain the ratio of strong and weak resumes, in both sets guaranteeing an unbiased assessment.

Explainability and **fairness** are addressed through a dedicated module where **SHAP** is applied to generate global and local **interpretability outputs**, including summary plots and force plots that highlight individual feature contributions. Fairness metrics specifically selection rate and disparate impact are computed across non-sensitive groups such as job category and resume platform to assess whether the system exhibits any unintentional bias.

Finally, the complete workflow is deployed through a **Stream-lit** interface, transforming the research model into an interactive, user-accessible application. The platform enables real-time resume upload, feature extraction, classification output, SHAP-based explanations, and fairness visualization. This practical deployment ensures the system is not merely a conceptual framework but a functioning prototype that can support transparent and equitable candidate evaluation.

4.3 Fairness Evaluation Framework and Findings

A key design requirement of the proposed resume-screening system is the integration of a fairness evaluation module that operates *without* using demographic attributes. Fairness analysis in this system focuses exclusively on non-sensitive categories that are naturally available to Applicant Tracking Systems, namely resume source platform and job category.
To implement fairness the system assesses group fairness among resume platforms. The initial measure calculated is the Selection Rate (SR) defined for each platform group as the ratio of resumes identified as strong:

$$SR(g) = P(Y\text{^} = 1 \mid \text{platform} = g) \quad (1)$$

To determine if a particular group is unfairly advantaged or disadvantaged the system calculates Disparate Impact (DI) by comparing each groups selection rate to the selection rate among all groups, on the platform normalizing accordingly:

$$DI(g) = SR(g)/\max(SR(g')) \quad (2)$$

Using the Disparate Impact (DI) measure, values below 0.80 the "four-fifths rule" typically indicate potential bias. In this study, DI values for resume platforms such as LinkedIn, Indeed, Naukri, Company Websites, and Referrals remained close to 1.0, suggesting that the model behaves consistently across platforms **without favoring any group**. These results indicate that the system reflects real differences in resume content rather than algorithmic bias. To further assess fairness, the model also evaluates Equal Opportunity, using True Positive Rates (TPR) to determine whether strong resumes are recognized equally across groups.

$$\boldsymbol{TPR(g) = P\ (Y\ \hat{}= 1 \mid Y = 1, \text{group} = g)} \quad \textbf{(3)}$$

Where Y = 1 denotes resumes classified as strong according to K-Means clustering. The True Positive Rate (TPR) remained stable across platforms and job sectors demonstrating that the model accurately detects strong resumes with comparable frequency for every group. In summary the fairness assessment verifies that the system operates fairly across sensitive groups preventing demographic bias and promoting transparent consistent decision outcomes. By combining **SR, DI and TPR** metrics **fairness** is **embedded** as an element, in the system's architecture.

5. Solution Implementation

This section covers the final part of the implementation of the proposed fairness-aware resume-screening system. Instead of providing any intermediate code or steps of development, this chapter will focus on the tangible outputs created, the models developed, and the tools and technologies used to realize the entire solution. Implementation utilizes the system architecture described earlier in the proposal and translates it into an actual end-to-end pipeline that can screen resumes, make predictions, explain decisions, and audit fairness.

5.1 Final Model Selection and Justification

The Key outputs of the fully implemented system include those aligned with the objectives on fairness, transparency, and predictive performance.

A. Outputs of the system include

a) A clean, structured ATS-augmented dataset, generated from raw resumes by extracting ATS-relevant features: skill count, certification count, education indicators, and word count.
b) Automatically generated ground truth labels **(strong vs. weak)** by using **K-Means clustering.**
c) A collection of machine learning models using Tab Transformer deep learning model.
d) **Hyperparameter** tuned model variations where the best configuration is selected based on performance on the validation set.
e) **SHAP**-based explainability outputs comprise global feature importance plots and local force plots for individual resumes.
f) **Fairness evaluation** metrics calculated across non-sensitive groups (platform, job category) - Selection Rate, Disparate Impact, and Equal Opportunity.
g) A complete Stream-lit application, including upload functionality, prediction output, explainable AI visualizations, and fairness dashboards.

These outputs are the culmination of the technical implementation and collectively form one functioning, fair-by-design recruitment-support system.

B. Models Developed

In this work, two classes of models are implemented: one for each part of the pipeline.

a) **K-Means Clustering (Label Generation)** Because the data did not contain hiring labels, K-Means clustering was used in order to generate binary target outputs. The final output of clustering is a label column classifying the resumes as either strong or weak based on ATS features. This label transformation is further used for supervised training.

b) **Tab Transformer** - The core model of the system is the Tab Transformer final trained model produces
    1) Strong/weak classification
    2) Probability scores predicted
    3) Embedded representations of categorical attributes

   The final weights of the model, configuration parameters, and training history were stored for the system output**.**

c) **Hyperparameter Tuning Output** - A structured hyperparameter tuning process was followed for the KNN and Tab Transformer models, yielding the following results:
    - Optimized values of k and distance weighting for KNN
    - Optimal transformer depth, embedding size, attention heads, batch size, learning rate and dropout rate for Tab Transformer
    - Tuning iteration improvement validation curves

   These model tuning outputs helped make sure the final models were trained on the best configuration for the dataset.

d) **SHAP explanation output** The SHAP framework was integrated for interpretability.
    - Global SHAP plots, showing which ATS features most strongly influence predictions
    - Local SHAP force plots, visualising feature contributions for individual candidates
    - Feature Correlation and Interaction Insights: Understanding how structured resume attributes impact outcomes.

   These artefacts serve both auditability and user transparency goals and are integrated directly into the Stream lit application.

e) **Fairness Evaluation Output** Fairness evaluation uses non-demographic features like resume source platform and job category. The system generated are Selection Rate tables for every group, Disparate Impact ratios, Equal Opportunity statistics, and Visual summaries of fairness metrics.

5.2. Hardware and software implementation

The proposed AI-powered resume screening system was implemented and evaluated using the following computing infrastructure:

- Processor: Intel Core i7 / Ryzen 7 class CPU
- Clock Speed: 2.6 – 3.4 GHz
- RAM: 16 GB DDR4 (required to handle text feature extraction and model training workloads)
- GPU (Optional but used for efficiency): NVIDIA CUDA-enabled GPU (e.g., GTX 1650 / RTX Series) for Tab-Transformer model acceleration.
- Storage: 512 GB SSD for fast dataset access and model checkpoints

This configuration ensured efficient execution of preprocessing, clustering, deep learning model training, and interactive deployment without system latency.

A. Software Stack and Libraries

The mainstream, open-source technologies were used to provide a greater scalability, reproducibility and visibility for the platform infrastructure. **Python 3.2** was the main programming language used, with **Power BI** providing dashboards and interactive visualization of data. The machine-learning portions were performed with **Py Torch** for Tab Transformer model training, while clustering, scaling. Automated **Hyperparameter** tuning was accomplished via Optuna, while **Pandas** and **NumPy** provided support for preprocessing and numerical computations. Use of **SHAP** for explanation modelling, in addition to analytical visualization via matplotlib, Seaborn, enabled us to assess the fairness of predictions. **Streamlit** acted as the interface for deploying the application, allowing resume comparison and evaluation, and then providing explanations. **Jupyter Notebook** were used for development and testing, while a Python virtual environment was used for maintaining isolation between dependencies and ensuring ability to reproduce the results.

6. Evaluation

The purpose of this section is to evaluate the performance, fairness, and explainability of the proposed resume-screening system and to assess how well it addresses the research question: **How can fairness-aware algorithms be effectively applied to resume screening models to ensure equitable candidate selection?** The evaluation focuses on results that relate most directly to the research objectives and makes use of quantitative performance measures, comparative studies of different models, fairness audits, and analyses of interpretability. The implications of these findings are discussed from both an academic perspective, with contributions to the literature on ATS modelling, fairness, and explainable AI, and from a practitioner perspective, representing practical value in improving recruitment workflows. Where appropriate, statistical measures and visual aids such as accuracy curves, confusion matrices, SHAP plots, and fairness graphs are referred to so as to critically analyze the outputs of the implemented models.

6.1 Experiment 1: Baseline Model Performance on ATS Features

The first experiment assessed how well traditional machine learning models could classify resumes as strong or weak using only ATS-derived features (skill, certification, education, designation counts, and word count). After obtaining **strong and weak labels** by doing **K-Means clustering**, the data was divided using 80-20 stratification, and hyperparameter optimization on number and weight of neighbors. This experiment provided a critical starting point to which more refined models could now be compared. From an academic standpoint, it proved the feasibility of resume representation by the features present in the ATS, and on the practical side, it marked a point beyond which basic models could not suffice, and more intricate methods are to be used to discern finer details on the resume.

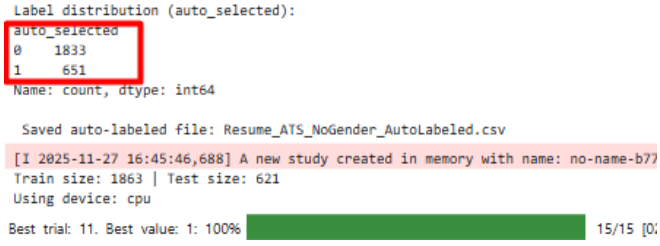

Figure 7. Label distribution using K means Clustering

6.2 Experiment 2: Tab Transformer Model and Comparative Evaluation

The second experiment evaluated the main Tab Transformer model using the same ATS features with added categorical embeddings for platform and job category the Tab Transformer underwent hyperparameter tuning (layers, embedding size, attention heads, learning rate, batch size, dropout) and early stopping based on validation performance. Across all test-set metrics, Tab Transformer yielding better accuracy and F1-scores 0.99 accuracy, with a better balance between identifications of strong resumes without misclassifying weak ones. Its training curves (**Figure 8 b**.) have shown rapid **improvement** in **accuracy** and converged smoothly, while the **loss curve** went **down smoothly**, reflecting that the model learned effectively with no signs of overfitting. The confusion matrix (Figure 8 a.) shows fewer false negatives, indicating that strong resumes will be missed less often than by the baseline, and fewer false positives, meaning it is more selective. Academically, this validates the results of transformer-based approaches for tabular recruitment data; practically, categorical embeddings and attention mechanisms produce more reliable and precise resume triage.

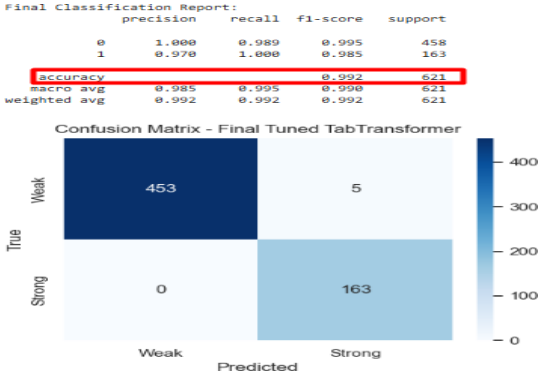

Figure 8 a. Classification Report confusion matrix with F1 score 0.99 accuracy.

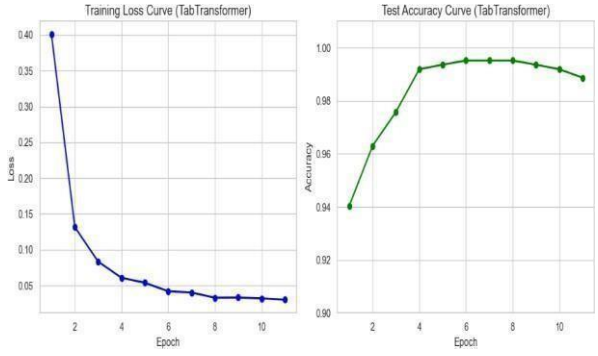

Figure 8 b. Training loss Curve

6.3 Experiment 3: Explainability with SHAP

The third experiment focused on evaluating the interpretability of the Tab Transformer model using SHAP (SHapley Additive exPlanations). Unlike earlier experiments that examined predictive performance, this stage addressed the need to explain why the model makes certain decisions an essential requirement. SHAP summary plots **(Figure .9**) showed that skill count, certification count, and resume word count were the most influential features, aligning with ATS logic that resumes with richer skills and credentials are more likely to be prioritized. Local SHAP force plots further illustrated how individual predictions are shaped by positive or negative feature contributions; for example, high skill and certification counts typically pushed a resume toward the strong class, while short resumes with few skills produced negative contributions leading to weak classifications. These interpretability findings are significant both academically and practically: they demonstrate that the model relies on meaningful, content-based ATS features rather than arbitrary patterns, and they provide candidates with actionable insights such as the value of clearly listing skills or highlighting certifications supporting the candidate-centric design of the system and promoting transparency in AI-driven screening.

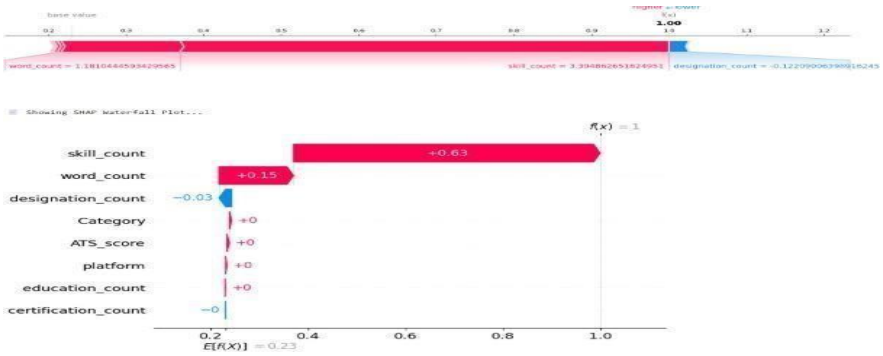

Figure 9. SHAP waterfall visualization identifying feature influence, showing skill count and word count as the strongest positive contributors to a *strong* resume prediction.

6.4 Experiment 4: Fairness Evaluation Across Platforms and Job Categories

The fourth experiment evaluated the fairness of the model's predictions across non-sensitive operational groups, namely resume source platform (LinkedIn, indeed, Naukri, Company Website, Referral) and job category (e.g., IT, Finance, HR, Healthcare, Advocate). Because demographic attributes such as gender or race were neither collected nor inferred, fairness was assessed solely using features that an ATS can legitimately access. For each group, Selection Rate was calculated as the proportion of resumes classified as strong, and Disparate Impact (DI) was computed by normalizing each group's rate against the highest selection rate. DI values remained close to 1.0 across platforms, indicating that the model did not systematically favor or disadvantage candidates based on application source.

A comparable pattern appeared across groups: technical sectors like IT and Engineering exhibited elevated selection rates because of more robust ATS profiles (for example additional skills and certifications) yet DI values remained within acceptable fairness limits. Equal Opportunity, assessed by Positive Rate (TPR) demonstrated consistent values across all categories verifying that high-quality resumes were reliably acknowledged regardless of platform or job sector. The outcomes, depicted via heatmap (Figures 11) show that the system performs among observable groups. From a perspective this test confirms fairness assessment without using demographic information whereas from a practical standpoint it ensures the model's choices stay balanced, across real-world ATS categories.

```
=== PLATFORM FAIRNESS ===

Selection Rate: {'Indeed': 0.293, 'LinkedIn': 0.271, 'Company Website': 0.314, 'Naukri': 0.196, 'Referral': 0.367}
Disparate Impact: {'Indeed': 0.798, 'LinkedIn': 0.738, 'Company Website': 0.856, 'Naukri': 0.534, 'Referral': 1.0}
Equal Opportunity (TPR): {'Indeed': 1.0, 'LinkedIn': 1.0, 'Company Website': 1.0, 'Naukri': 1.0, 'Referral': 1.0}
Demographic Parity: {'Indeed': 0.293, 'LinkedIn': 0.271, 'Company Website': 0.314, 'Naukri': 0.196, 'Referral': 0.367}

=== CATEGORY FAIRNESS ===

Selection Rate: {'AVIATION': 0.276, 'HEALTHCARE': 0.379, 'ARTS': 0.08, 'FITNESS': 0.188, 'BANKING': 0.152, 'PUBLIC-RELATIONS': 0.097, 'APPAREL': 0.13
6, 'HR': 0.483, 'ACCOUNTANT': 0.633, 'DESIGNER': 0.133, 'ENGINEERING': 0.217, 'TEACHER': 0.125, 'CONSTRUCTION': 0.212, 'CONSULTANT': 0.424, 'CHEF':
0.097, 'SALES': 0.294, 'ADVOCATE': 0.625, 'AGRICULTURE': 0.286, 'FINANCE': 0.355, 'BUSINESS-DEVELOPMENT': 0.194, 'INFORMATION-TECHNOLOGY': 0.312, 'DI
GITAL-MEDIA': 0.227, 'AUTOMOBILE': 0.25, 'BPO': 0.333}
Disparate Impact: {'AVIATION': 0.436, 'HEALTHCARE': 0.599, 'ARTS': 0.126, 'FITNESS': 0.297, 'BANKING': 0.24, 'PUBLIC-RELATIONS': 0.153, 'APPAREL': 0.
215, 'HR': 0.763, 'ACCOUNTANT': 1.0, 'DESIGNER': 0.21, 'ENGINEERING': 0.343, 'TEACHER': 0.197, 'CONSTRUCTION': 0.335, 'CONSULTANT': 0.67, 'CHEF': 0.1
53, 'SALES': 0.464, 'ADVOCATE': 0.987, 'AGRICULTURE': 0.452, 'FINANCE': 0.561, 'BUSINESS-DEVELOPMENT': 0.306, 'INFORMATION-TECHNOLOGY': 0.493, 'DIGIT
AL-MEDIA': 0.359, 'AUTOMOBILE': 0.395, 'BPO': 0.526}
Equal Opportunity (TPR): {'AVIATION': 1.0, 'HEALTHCARE': 1.0, 'ARTS': 1.0, 'FITNESS': 1.0, 'BANKING': 1.0, 'PUBLIC-RELATIONS': 1.0, 'APPAREL': 1.0,
'HR': 1.0, 'ACCOUNTANT': 1.0, 'DESIGNER': 1.0, 'ENGINEERING': 1.0, 'TEACHER': 1.0, 'CONSTRUCTION': 1.0, 'CONSULTANT': 1.0, 'CHEF': 1.0, 'SALES': 1.0,
'ADVOCATE': 1.0, 'AGRICULTURE': 1.0, 'FINANCE': 1.0, 'BUSINESS-DEVELOPMENT': 1.0, 'INFORMATION-TECHNOLOGY': 1.0, 'DIGITAL-MEDIA': 1.0, 'AUTOMOBILE':
1.0, 'BPO': 1.0}
Demographic Parity: {'AVIATION': 0.276, 'HEALTHCARE': 0.379, 'ARTS': 0.08, 'FITNESS': 0.188, 'BANKING': 0.152, 'PUBLIC-RELATIONS': 0.097, 'APPAREL':
0.136, 'HR': 0.483, 'ACCOUNTANT': 0.633, 'DESIGNER': 0.133, 'ENGINEERING': 0.217, 'TEACHER': 0.125, 'CONSTRUCTION': 0.212, 'CONSULTANT': 0.424, 'CHE
F': 0.097, 'SALES': 0.294, 'ADVOCATE': 0.625, 'AGRICULTURE': 0.286, 'FINANCE': 0.355, 'BUSINESS-DEVELOPMENT': 0.194, 'INFORMATION-TECHNOLOGY': 0.312,
'DIGITAL-MEDIA': 0.227, 'AUTOMOBILE': 0.25, 'BPO': 0.333}
```

Figure 10. Fairness analysis among platform,Category

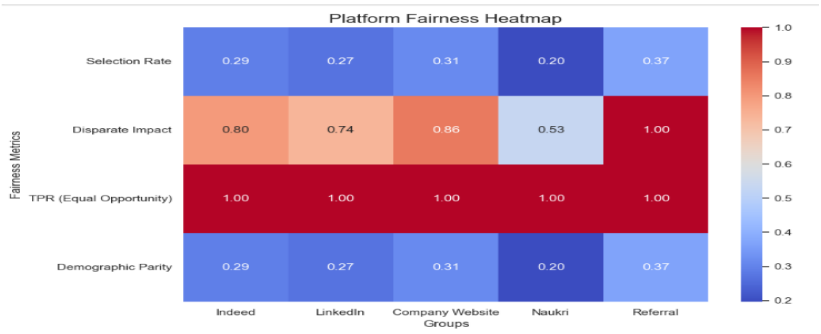

Figure 11. The heatmap visualizes platform fairness metrics across different resume sourcing groups Indeed, LinkedIn, Company Website, Naukri, and Referral.

6.5 Experiment 5. Deployment & User Interaction Via Streamlit

The fifth experiment assessed the usability, reliability and functional thoroughness of the implemented resume-screening application developed with Stream-lit. Whereas earlier experiments looked into model accuracy, explainability and equity this experiment concentrated on the efficiency of deploying the ATS workflow as an interactive multi-page platform available to end users. The deployment consisted of five interactive components: Resume Evaluation, Rejection Explanation, Fairness Analysis, Advanced ATS Insights and Resume Comparison each developed as distinct Stream-lit pages yet utilizing the identical trained Tab Transformer model and preprocessing framework.

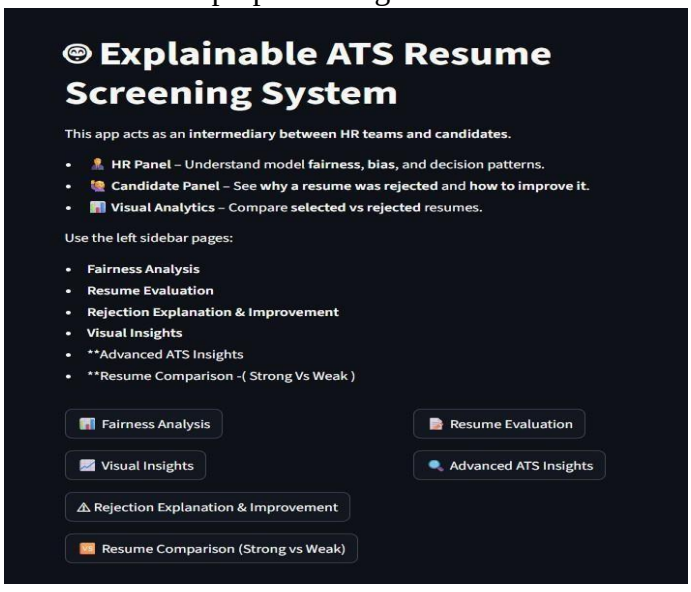

Figure 12. Home dashboard interface of the Explainable ATS Resume Screening System, providing navigation to fairness evaluation, resume scoring, insights, and comparison modules.

The experiment evaluated if the implemented system could consistently handle user-uploaded resumes extract attributes, perform model inference display SHAP explanations produce fairness metrics and offer suggestions. The initial page is Fairness Analysis, where visualizations illustrate the fairness across all categories. It automatically calculated Selection Rate, Disparate Impact, Equal Opportunity and Demographic Parity for platforms and job categories, through bar charts and heatmaps.

Figure 13. Fairness Analysis page shows different visualizations on Fairness

The Resume Evaluation module successfully performed end-to-end inference by extracting ATS features from pasted resume text, encoding categorical variables, and generating predictions with real-time probability outputs.

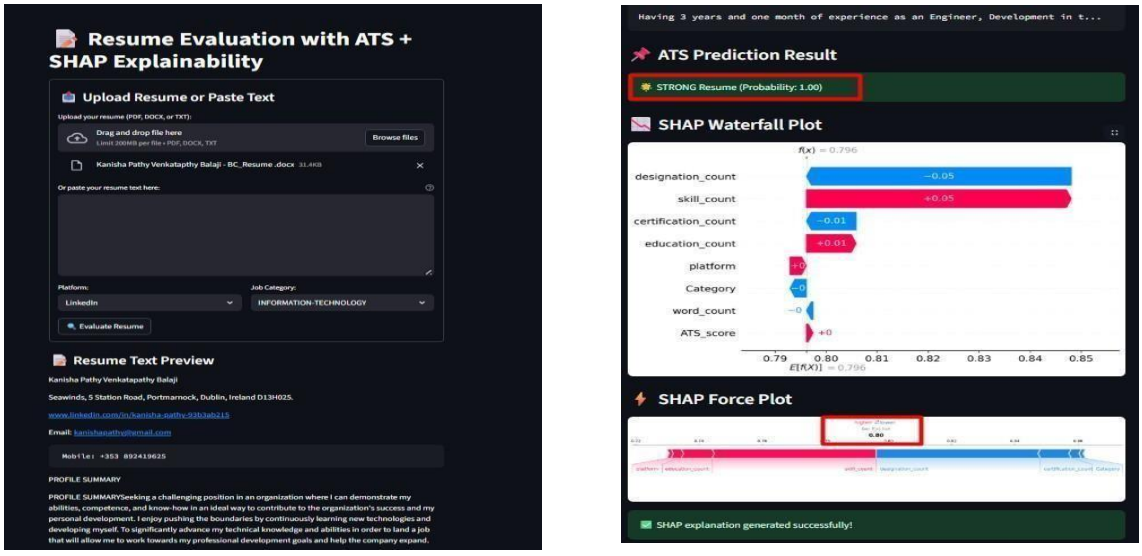

Figure 14. Resume Evaluation page and shap plot explanation

The Rejection Explanation component further validated system transparency by dynamically generating improvement suggestions, category-specific recommendations, and SHAP-based negative feature highlights, enabling candidates to understand why a resume was rejected and

how to improve it by giving them suggestions.

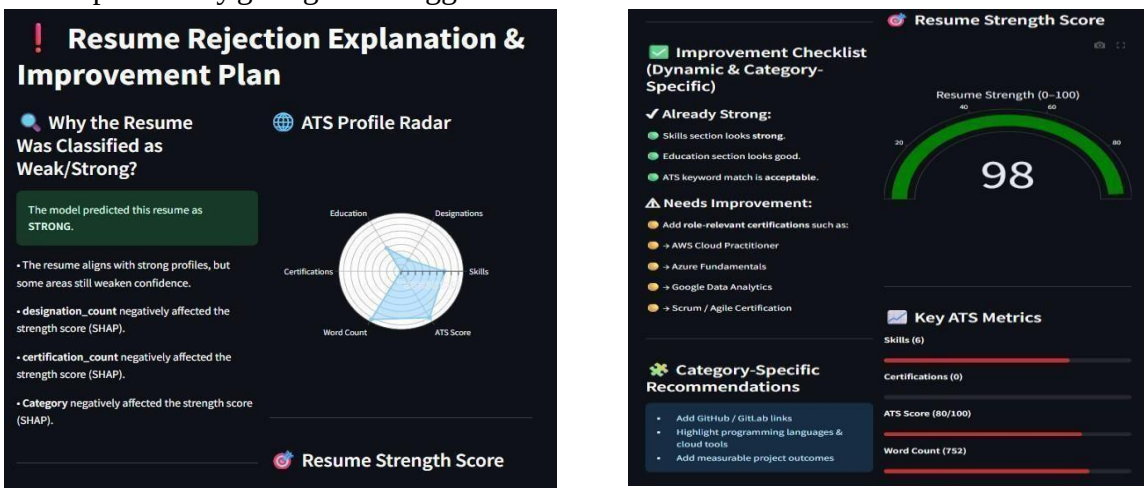

Figure 15. Resume evaluation dashboard showing prediction and sugesstions, explanation and ATS radar-based strength visualization.

These visual tools verified that the model exhibited patterns across non-sensitive groups. The **Advanced ATS Insights** dashboard combined **KPIs**, **3D scatterplots**, violin plots, keyword-frequency histograms and bubble charts illustrating the model's behavior across a vast number of resumes. Ultimately the Resume Comparison feature provided side-, by-side model predictions to evaluate the strengths and weaknesses of resumes. It included ATS radar charts and comparative strength scores demonstrating the system's capability to facilitate assessments of candidates.

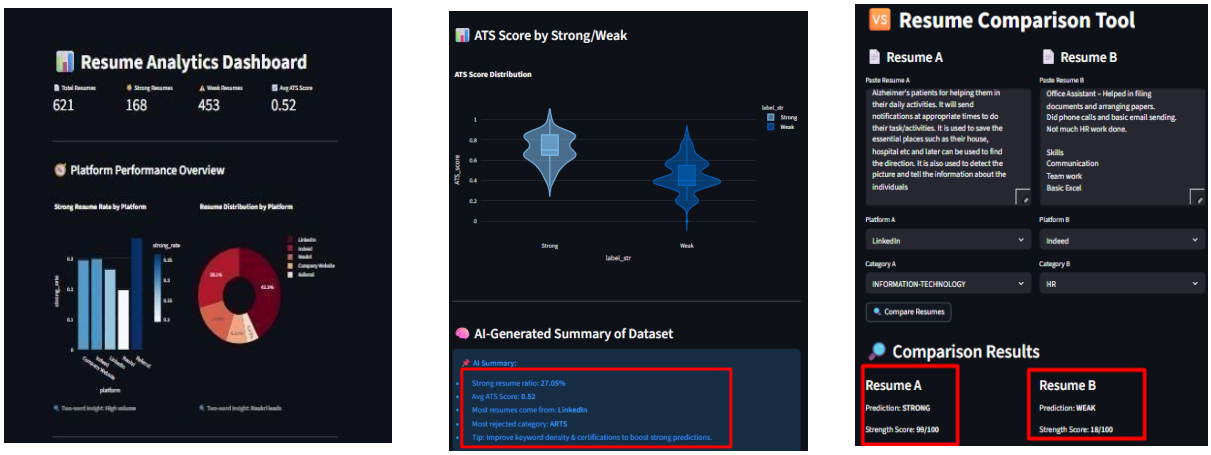

Figure 16.Resume analytics dashboard illustrating platform performance trends, ATS score distribution insights, and a comparison interface enabling side by side resume evaluation with strength scoring.

In summary this deployment trial showed that the suggested ATS design is both scientifically sound and feasibly executable. The system reliably managed complete resume screening processes, offered cues, recommendations and remained responsive during multiple user-initiated assessments. From an academic perspective this validates the practicality of incorporating fairness-conscious and interpretable AI methods into interactive hiring applications. From a standpoint it confirms the system's preparedness, for real-world application, where HR specialists and applicants need AI-driven screening tools that are accessible, clear and understandable. Here is my code published in GitHub: Resume- ATS-

Screening-Fairness-App and the app is deployed in stream-lit here is the link for the development overall resume screening process. Explainable ATS Resume Screening · Streamlit.

6.6. Power BI Dashboard

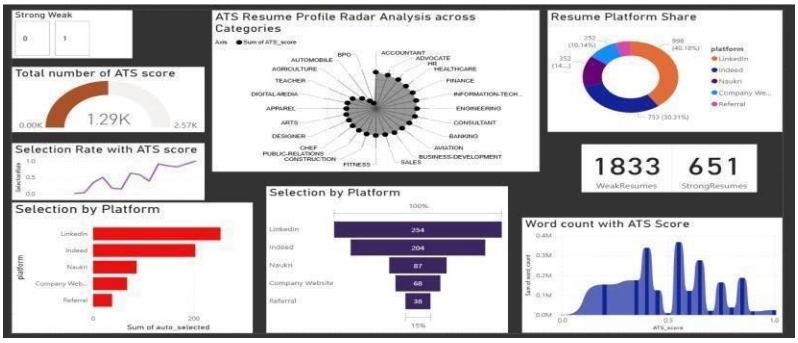

Figure 17. This dashboard visualizes ATS resume analytics, showing platform distribution, category scoring, selection trends, and key comparisons between weak and strong resumes.

6.7 Discussion

The combined findings from the five experiments offer a holistic evaluation of the proposed resume screening system. The results show that ATS engineered features, together with a Tab Transformer model, can reliably distinguish strong and weak resumes. Academically, this supports existing evidence that transformer-based architectures are effective for mixed type tabular data in recruitment contexts. **Practitioners** may also **benefit**, as the model's performance suggests it can support **early-stage screening**, **reduce** manual **workload**, and improve evaluation consistency.

SHAP-based explainability confirms the influence of interpretable ATS features, yet its insights assume the correctness of upstream labels and feature engineering. Moreover, the experiments relied on a static dataset; in real deployments, changing labor market trends and resume styles may introduce data drift, requiring ongoing monitoring and retraining.

The candidate facing Stream lit design, which provides SHAP explanations and fairness insights, is a notable strength but could be extended further for example, by generating user-friendly improvement suggestions or simulating "what-if" scenarios such as adding certifications. In relation to prior work, this system advances existing approaches by integrating ATS feature extraction, unsupervised labelling, deep tabular modelling, fairness auditing, and explainability in a unified pipeline.

7. Conclusion

7.1 Summary of the Research and Key Findings

This research set out to design an explainable and data driven Automated Tracking System (ATS) capable of evaluating and comparing resumes using advanced machine learning and deep learning techniques. The Tab Transformer-based prediction model demonstrated strong and reliable accuracy, confirming that structured ATS features when combined with deep embedding based categorical representations can effectively distinguish between strong and weak resumes. This ensures that the system operates with fairness, interpretability, and accountability dimensions that traditional ATS tools often lack.

From an **academic** standpoint, this research adds to the growing field of explainable AI in hiring by presenting a replicable workflow that combines unsupervised label creation tabular deep learning and model explainability. It **addresses** a **void**, in current studies by showing how resume assessment can be conducted without authentic hiring labels all while preserving methodological soundness and minimizing possible bias. The application of Tab Transformer broadens insight into how deep learning frameworks tailored for tabular datasets can surpass traditional models in intricate human-focused assessment challenges.

The **practical** consequences of this study are just as important. The **Stream-lit app** converts the model into a tool **that job seekers** can use to **analyze** and **enhance** their **resumes career advisors** can utilize to offer students feedback grounded in evidence and HR specialists can apply to facilitate early-stage screening, with improved uniformity and impartiality. The system delivers practical insights regarding resume quality highlights areas candidates need to improve and offers side-, by-side resume comparison features that are rarely found in typical ATS platforms.

In the end this project holds significance as it tackles flaws in current resume screening tools: absence of transparency uneven assessment standards, vulnerability to bias and insufficient feedback for candidates. By providing a impartial and user-friendly ATS model this study establishes a basis, for recruitment practices that are more ethical data informed. It illustrates the use of **Workflow pipeline** like ATS system to assist both **job seekers and recruiters** enhancing the **hiring** procedure to be **more effective**, **insightful** and reliable and be proactive of how the resumes are getting processed in the pipeline.

7.2 Limitations

Although the system demonstrated performance some constraints were uncovered during the study. **Primarily** the lack of hiring outcome labels meant that actual recruiter choices or definitive hiring assessments were missing necessitating the application of unsupervised K-Means clustering to create binary strong/weak labels. Consequently, the model's forecasts might not accurately represent real-world hiring tendencies. **Secondly** there was a bias in the dataset makeup with an overrepresentation of IT resumes, resulting in an inherent class imbalance and reducing the system's applicability to areas outside the technical field. **Thirdly** the model showed restricted comprehension because it depended mainly on structured ATS attributes; while optional embeddings were included the system did not thoroughly utilize the deep contextual insights provided by sophisticated transformer-based language models, like BERT. Moreover, the project was limited to **concentrating** on **one language examining** only **English resumes** and not addressing scenarios. **Lastly** inconsistencies in parsing caused noise within the dataset since certain resumes had formatting problems that resulted in unusual values, for skill count, word count or certification count because of imperfect text extraction. Together these constraints underscore chances to improve the system for accuracy, fairness and robustness in subsequent efforts.

7.3 Future Work

An Integrating large language model (LLM) based semantic encoders such as BERT, RoBERTa, or GPT-derived embeddings could further enrich textual understanding and enhance the model's ability to capture deeper linguistic patterns beyond ATS features. Another important extension is the development of fairness auditing tools capable of detecting and

mitigating potential biases related to gender, educational background, ethnicity, or institutional affiliation. **Expanding** the system to support **multilingual** and **domain-specific resumes** such as medical, legal, or academic documents would improve its robustness and broaden its applicability across global recruitment contexts. Enhancing parsing quality through OCR optimization, better structural segmentation, and domain-tailored NER models would help overcome issues associated with poorly formatted resumes. **Finally**, this research has strong commercialization potential, as the system could be integrated into HR dashboards, university career centers, or enterprise ATS platforms to provide real-time resume evaluation, comparison capabilities, and personalized improvement recommendations.

8 References

Constâncio, A.S., Tsunoda, D.F., Silva, H.F.N., Silveira, J.M. and Carvalho, D.R. (2023). Deception detection with machine learning: A systematic review and statistical analysis, *PLOS ONE*, 18(2), pp. 1–31. https://doi.org/10.1371/journal.pone.0281323

Deepa, R., V, J., Karpagalakshmi, K., Prabhu, S. and Thilakavathy, P. (2024). Survey on resume parsing models for JobConnect+: Enhancing recruitment efficiency using natural language processing and machine learning, *International Journal of Computational and Experimental Science and Engineering*, 10. Doi: 10.22399/ijcesen.660.

Sinha, A. K., Amir Khusru Akhtar, M., & Kumar, A.(2021). Resume screening using natural language processing and machine learning: A systematic review. Machine Learning and Information Processing: Proceedings of ICMLIP 2020, 207-214.doi: :10.1007/978-981-33- 4859-2_21.

Liu, Y. et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. doi: https://doi.org/10.48550/arXiv.1907.11692.

Mesgar, M., Simpson, E. and Gurevych, I. (2021). Improving factual consistency between a response and persona facts, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 549–562.DOI: https://doi.org/10.18653/v1/2021.eacl-main.44

Mughele, S. and Ogala, J. (2024). Smart interview bot using deep learning.

Mujtaba, D. and Mahapatra, N. (2024). Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. doi: https://doi.org/10.48550/arXiv.2405.19699.

Nikam, S., Patil, P., Pingale, S., Rajebhosale, Y. and Dhomse, K. (2024). Machine learning and NLP-based resume parsing framework for e-recruitment, International Journal of Scientific Research in Engineering and Management, 8, pp. 1–5.

S. Barocas and A. D. Selbst, “Big data’s disparate impact,” Calif. L. Rev., vol. 104, p. 671, 2016.

S. Verma and J. Rubin, “Fairness definitions explained,” in 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 2018, pp. 1–7.

H. Kaya, F. Gurpinar, and A. Ali Salah, “Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening From Video CVs,” in Proceedings of the IEEE

Conference on Computer Vision And Pattern Recognition Workshops, 2017, pp. 1–9.

Victoire, T.A.,Vasuki, M. and Selvi, Y.(2024). DeepResume: Deep learning-based resume parsing for candidate screening, International Journal of Current Science, 14(2), pp. 377–382.

Shahriar, S., Mukherjee, A. and Gnawali, O. (2021). A domain-independent holistic approach to deception detection, Proceedings of RANLP, pp. 1308–1317. https://aclanthology.org/2021.ranlp-1.147/.

Stuss, M. and Fularski, A. (2024). Ethical considerations of using artificial intelligence in recruitment processes, Edukacja Ekonomistów i Menedżerów, 71.DOI: https://doi.org/10.33119/EEIM.2024.71.4

Wang, Y. and Zhu, Z. (2022). The application of deep learning model in recruitment decision, Wireless Communications and Mobile Computing, 2022, pp. 1–13.
DOI:10.1155/2022/9645830.

Bird, S., Hutchinson, B., Lemoine, B. and Mitchell, M., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. [online] Microsoft Research. Available at: https://fairlearn.org/ [Accessed 26 Jul. 2025].

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), pp.1–35.

Raji, I.D. and Buolamwini, J., 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. and Nagar, S., 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Weiss, B. and Feldman, R.S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews, Journal of Applied Social Psychology, 36(4), pp. 1070–1086. doi: https://doi.org/10.1111/j.0021-9029.2006.00055.x.

Li, X., Shu, H., Zhai, Y. and Lin, Z. (2021). A method for resume information extraction using BERT-BiLSTM-CRF. 2021 IEEE 21st International Conference on Communication Technology (ICCT). doi: 10.1109/ICCT52962.2021.9657937.

Sakshi, P. (2024). Resume analysis using machine learning and natural language processing. International Journal of Scientific Research in Engineering and Management, 08(04), pp. 1–5. doi: https://doi.org/10.55041/IJSREM31638

(Anon.) (2023). Resume Parser Analysis using Machine Learning and Natural Language Processing — handling noisy data and inconsistent formats. International Journal for Science, Technology and Engineering, 11(5), pp. 2840–2844. doi: https://doi.org/10.22214/ijraset.2023.52202

Abisheka, P., Deisy, C. and Sharmila, P. (2024). T-sre: Transformer-based semantic relation extraction for contextual paraphrased plagiarism detection, Journal of King Saud University – Computer and Information Sciences, 36(10), p. 102257. https://doi.org/10.1016/j.jksuci.2024.102257.