**ISSN:** 2583-9888
**DOI:** doi.org/10.51219/JAIMLD/Yusuf-Jazakallah/16

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 4

*Research Article*

# Effective Strategies for Mitigating Bias in Hiring Algorithms: A Comparative Analysis

**Yusuf Jazakallah***

Recruitment Smart Technologies, London, UK

**Citation:** Jazakallah Y. Effective Strategies for Mitigating Bias in Hiring Algorithms: A Comparative Analysis. *J Artif Intell Mach Learn & Data Sci*, 1(4), 125-134. DOI: doi.org/10.51219/JAIMLD/Yusuf-Jazakallah/16

**Received:** 29 August, 2023; **Accepted:** 26 September, 2023; **Published:** 16 October, 2023

***Corresponding author:** Mr. Yusuf Jazakallah, Recruitment Smart Technologies, London, UK. Email: yusuf@recruitmentsmart.com



## ABSTRACT

Bias in hiring algorithms is a critical issue that has been widely recognized in recent years. As more companies rely on automated candidate selection processes, it is essential to develop fair and equitable recruitment practices that ensure equal opportunities for all candidates. The objective of this research paper is to propose a comprehensive framework for mitigating bias in hiring algorithms. By utilizing a combination of machine learning techniques, statistical analysis, and ethical considerations, the study aims to identify, measure, and mitigate both overt and subtle forms of bias present in these algorithms. This paper's findings underscore the significance of employing de-biasing strategies to ensure diversity and inclusion in the workplace. In this introduction, we will discuss the critical issue of bias mitigation in hiring algorithms, the importance of fair and equitable recruitment practices, and the objective of the study. We will also provide an overview of the research methodology, the measurement of bias, and the proposed mitigation strategies. Finally, we will summarize the key findings and the proposed framework for reducing bias in hiring algorithms.

**Keywords:** Bias mitigation; Hiring algorithms; Fairness; Diversity; Inclusion; Ethics; Machine learning

## 1 Introduction To Bias Mitigation In Hiring Algorithms

### 1.1 What is the critical issue of bias mitigation in hiring algorithms?

Automated hiring systems (AHSs) are being used in the hiring process and are claimed to detect and mitigate discrimination against protected groups[1]. 98% of the Fortune 500 companies have adopted Applicant Tracking Systems of some kind[1]. However, there is a lack of transparency and it is feared that access to jobs for specific profiles may be limited[1]. The UK has a different legal background than the US in terms of hiring, equality law, and data protection law[1], and this could be important for addressing concerns about transparency. Bias mitigation is a critical issue in hiring algorithms, as AI has the potential to shape the future of work and the workforce[1]. Machine learning classification may introduce or reinforce bias in the hiring process[2] and it is important to ensure that diverse groups are represented in the development and testing of hiring algorithms[1]. Empirical analysis has indicated that bias mitigation is a feasible approach to ensure group fairness in hiring algorithms; however this comes at a cost of efficiency and accuracy[2]. It is, therefore, essential to ensure transparency in hiring algorithms to detect and address biases[1]. Despite the various claims of 'bias mitigation' in AHSs, claims are rarely scrutinized and evaluated, making bias mitigation a critical issue that needs to be addressed[1].

### 1.2 Why is it important to develop fair and equitable recruitment practices?

It is increasingly important to address the potential for unfairness in recruitment practices in order to promote equity and access[3]. While algorithmic decision-making in recruitment can lead to discrimination and unfairness[4], relational equality in recruitment practices has been shown to be a promising approach for promoting fairness and equity[3]. On the other hand, distributional equality in recruitment practices can fail to promote fairness and equity[3]. Therefore, it is essential to develop fair and equitable recruitment practices to help address these issues and promote greater diversity and inclusion in the workforce[3]. Fair and equitable recruitment practices are important because

discrimination in hiring is a persistent problem in many sectors, and it can lead to negative consequences for those who are discriminated against, including reduced opportunities for employment and lower wages[3]. Companies offering algorithms for employment assessment need to disclose their development and validation procedures in order to evaluate their practices[3]. Algorithmic de-biasing techniques pose challenges for antidiscrimination law, and technical and legal perspectives need to be considered to develop fair and equitable recruitment practices[3]. The use of algorithms in hiring has grown rapidly, but little is known about how these methods are used in practice[3]. Fair and equitable recruitment practices are important to address or mitigate bias in hiring[3]. Ethical considerations should guide the development and implementation of AI-enabled recruitment practices[4]. Developing fair and equitable recruitment practices can lead to a diverse and inclusive workforce, whereas the use of AI in recruitment can lead to bias and discrimination[4]. Unfair recruitment practices can result in negative consequences for both individuals and organizations[4], and discrimination in recruitment can lead to unequal opportunities for job seekers[3]. Therefore, it is important to develop fair and equitable recruitment practices to help reduce discrimination in the labor market[3].

### 1.3 What is the objective of the study?

The objective of the present study is to explore the potential of talent acquisition software for fostering equity in the hiring process for underrepresented professionals[5]. To this end, the study offers a critical analysis of talent acquisition software, using equity as a central concept[5]. This will foster a richer understanding of what fairness means, and consider algorithmic bias in talent acquisition[5]. Moreover, feminist design thinking is used as a framework for evaluating how AI software might impact marginalized populations[5]. The study also aims to explain and provide a guide on the use of the Stanford revision and extension of the Binet-Simon intelligence scale[3], as well as present feminist design thinking as a theoretical lens for mitigating algorithmic bias[5]. Additionally, the study examines sources of algorithmic bias in talent acquisition software[5], and challenges of enforcing these laws in practice, particularly due to the opacity of automated hiring[6]. Furthermore, the study seeks to identify the discriminatory effects of these algorithms for legally protected groups, and to find a balance between the harms and benefits of using these tools, as outlined in equality and data protection laws[6]. Lastly, the study evaluates the application of UK law to the use of complex algorithms in job application assessments, and suggests the introduction of a transparent recruitment scheme to incentivize the publication of equality metrics contained in employers' data protection impact assessments[6]. This scheme should be a collaborative effort between the Information Commissioner's Office and the Equality and Human Rights Commission[6].

## 2. Research Methodology

### 2.1 What techniques are used to identify, measure, and mitigate bias in hiring algorithms?

The research methodology employed in this study is designed to comprehensively address bias mitigation in hiring algorithms by leveraging a triad of techniques: machine learning, statistical analysis, and ethical considerations. The overarching goal of the methodology is to uncover, quantify, and mitigate instances of bias within hiring algorithms, encompassing both overt and nuanced manifestations of bias.

#### 2.1.1 Machine Learning Techniques

The foundation of our approach lies in the utilization of machine learning techniques to analyze and model candidate selection patterns. We begin by curating a large dataset of historical hiring decisions, including candidate profiles and outcomes. Through this dataset, we implement state-of-the-art machine learning algorithms, such as supervised learning classifiers and clustering methods, to identify patterns and relationships within the data. To address overt biases, the methodology employs techniques like re-weighting and adversarial training. Re-weighting assigns appropriate weights to different subgroups within the dataset to counteract overrepresentation or underrepresentation. Adversarial training aims to minimize the distinguishability of sensitive attributes (e.g., gender, ethnicity) within the learned features, thereby promoting fairness.

#### 2.1.2. Statistical Analysis

Statistical analysis is an essential component to quantitatively assess the impact of biases in the hiring algorithm. We conduct detailed statistical examinations, including regression analysis, propensity score matching, and A/B testing, to quantify the association between algorithmic decisions and candidate attributes. These analyses provide insights into bias magnitudes, directionality, and potential causal relationships. By examining group-based disparities in hiring outcomes, we aim to uncover subtle biases that might not be immediately evident. Statistical analysis enables us to pinpoint specific stages in the recruitment process where bias is most prominent, facilitating targeted interventions.

#### 2.1.3. Ethical Considerations

The integration of ethical considerations is a core aspect of our research methodology. We engage in ongoing dialogues with experts in AI ethics, organizational psychology, and diversity and inclusion to ensure that the research is guided by ethical principles. Furthermore, we actively involve stakeholders from diverse backgrounds to provide input on potential biases and their implications. Ethical considerations extend beyond algorithm design and encompass the broader context of bias within the hiring ecosystem. Our methodology promotes transparency, accountability, and fairness by involving diverse perspectives and incorporating feedback loops that allow for continuous refinement of the algorithm.

### 2.2 How are machine learning, statistical analysis, and ethical considerations employed in the research methodology?

#### 2.2.1 Machine Learning Techniques

Machine learning techniques form a cornerstone of the research methodology, providing the tools to analyze, model, and address bias in hiring algorithms. These techniques encompass a spectrum of approaches designed to uncover and rectify biases at various stages of the candidate selection process.

**a. Data Preprocessing and Feature Engineering:**

The research begins with thorough data preprocessing and feature engineering. Raw candidate data is cleaned, standardized, and transformed into informative features. Special attention is given to features that are potentially sensitive or prone to bias, such as gender, ethnicity, and educational background. Careful consideration is taken to ensure that sensitive attributes are treated appropriately to avoid amplifying biases during subsequent modeling.

**b. Supervised Learning for Bias Identification:**

Supervised learning techniques are employed to develop predictive models that capture the hiring algorithm's decision-making process. These models are trained on historical hiring data, learning to predict whether a candidate will be selected or rejected based on their attributes. By comparing model predictions with actual outcomes, discrepancies can be identified. Statistical analyses, such as confusion matrix metrics and fairness-aware evaluation metrics (e.g., disparate impact, equal opportunity), are used to quantitatively measure the presence and extent of bias.

**c. De-biasing Techniques:**

De-biasing methods are pivotal in mitigating biases in hiring algorithms. Two primary approaches are adopted:

**Re-weighting:**

Biased training data may lead to model bias. Re-weighting assigns higher weights to underrepresented groups and lower weights to overrepresented groups, effectively balancing the dataset and reducing the impact of biased training instances.

**Adversarial Training:**

Adversarial training introduces a separate neural network (adversary) tasked with distinguishing between sensitive attributes within the model's learned features. The main model aims to minimize the adversary's ability to distinguish these attributes, resulting in learned features that are less sensitive to bias.

**d. Fair Representation Learning:**

Incorporating fairness into the representation learning process is another vital aspect of bias mitigation. Fair representation learning techniques, such as adversarial de-biasing and adversarial re-ranking, are employed. These methods work to transform the learned feature space, disentangling sensitive attributes from non-sensitive ones, ultimately producing more equitable and unbiased representations.

**e. Model Evaluation and Iteration:**

The trained models are rigorously evaluated using fairness-aware metrics, accuracy, and other relevant evaluation criteria. The iterative process involves analyzing the model's behavior, identifying areas of bias propagation, and fine-tuning the algorithms accordingly. Continuous monitoring of model performance and bias mitigation effectiveness ensures that the developed models align with the desired fairness goals.

**2.2.2 Statistical Techniques**

Statistical analysis forms a crucial component of the research methodology, providing the means to quantify bias and assess its impact on hiring algorithms. By employing various statistical techniques, the research aims to uncover hidden biases and their implications in candidate selection processes.

**a. Regression Analysis:**

Regression analysis is employed to investigate the relationship between candidate attributes and hiring outcomes. Multiple regression models are built, accounting for various candidate characteristics, such as educational background, experience, and demographic information. This analysis helps quantify the influence of different attributes on hiring decisions, thereby revealing any potential biases associated with specific attributes.

**b. Propensity Score Matching:**

Propensity score matching is used to address selection bias and assess the effect of candidate attributes on hiring outcomes. By matching candidates with similar propensity scores across different demographic groups, the analysis controls for confounding variables and isolates the impact of specific attributes on candidate selection. This technique helps identify disparities in hiring rates among different demographic groups while accounting for other factors.

**c. A/B Testing:**

A/B testing, commonly used in experimental design, is adapted to evaluate the effectiveness of bias mitigation interventions. Controlled experiments are conducted wherein different versions of the hiring algorithm are tested. One version incorporates bias mitigation techniques, while the other serves as a control. A/B testing enables the quantification of bias reduction and provides insights into the real-world impact of bias mitigation strategies.

**d. Group-Based Disparities Analysis:**

Group-based disparities analysis focuses on assessing hiring outcomes across different demographic groups. Statistical techniques, such as chi-square tests and t-tests, are employed to identify statistically significant differences in selection rates, interview invitations, and other relevant metrics. This analysis helps identify both overt and subtle biases that might disproportionately affect certain groups.

**e. Impact Assessment:**

The impact of bias mitigation interventions is assessed using fairness-aware evaluation metrics. These metrics, including disparate impact, equal opportunity, and demographic parity, provide quantifiable measures of bias reduction. By comparing the results before and after applying bias mitigation strategies, the research assesses the extent to which bias has been mitigated within the hiring algorithm.

**2.2.3 Ethical Considerations**

Ethical considerations serve as a fundamental pillar of the research methodology, guiding the approach to bias mitigation in hiring algorithms. Incorporating ethical principles ensures that the research not only identifies and mitigates biases but also upholds fairness, transparency, and inclusivity throughout the process.

**a. Collaborative Ethics Framework:**

The research actively engages with experts in AI ethics, organizational psychology, diversity and inclusion, and related fields. Collaborative discussions involving interdisciplinary stakeholders foster a nuanced understanding of the ethical challenges associated with bias in hiring algorithms. Insights from these discussions guide the development of the research framework and shape the application of bias mitigation strategies.

**b. Informed Consent and Data Privacy:**

Ethical considerations encompass obtaining informed consent from all parties involved. Data subjects, such as candidates and hiring managers, are informed about the research purpose, data usage, and potential implications of the study. Additionally, stringent data privacy protocols are implemented to safeguard sensitive candidate information, adhering to legal and ethical standards.

**c. Transparency and Algorithm Explainability:**

Ensuring transparency in algorithmic decision-making is a central ethical consideration. The research focuses on developing algorithms that are interpretable and explainable, allowing candidates and stakeholders to understand the factors influencing selection outcomes. Transparent algorithms empower candidates to make informed decisions and hold organizations accountable for their hiring practices.

**d. Feedback Loops and Continuous Improvement:**

Ethical considerations extend beyond the research phase and into the implementation of bias mitigation strategies. The research promotes the establishment of feedback loops that enable ongoing refinement of the algorithms. Feedback from candidates, hiring managers, and other stakeholders helps identify potential issues, unintended consequences, and opportunities for improvement, ensuring that biases are actively addressed over time.

**e. Fairness and Inclusivity:**

The ethical underpinning of the research methodology emphasizes fairness and inclusivity. The algorithms are designed not only to mitigate bias but also to enhance diversity in candidate selection. Ethical considerations guide the development of strategies that promote equal opportunities, encourage diverse talent pools, and contribute to the creation of inclusive work environments.

### 2.3 What is the significance of de-biasing strategies in recruitment automation systems?

De-bias strategies are essential in the development of a fair and equitable recruitment process[15]. Research methodology is a key factor to consider when designing a de-bias strategy[7,8,9]. It involves studying the methods used in the field, determining the nature of the study, and establishing the purpose and research design. Research methods, such as surveys or interviews, are the tools used to gather data, while research methodology is the set of procedures used to identify, select, process, and analyze the information[11]. This includes how the researcher intends to tackle issues like collection methods, data analysis and interpretation[14]. It is important to consider the technical and legal perspectives when developing a research methodology, as incorrect choices can lead to low quality research[8]. The correct choice of research methodology is essential to ensure that the results are fair and equitable[8]. By applying de-bias strategies to recruitment automation systems, organizations can ensure that they are creating a diverse and inclusive workforce.

## 3. Measurement of Bias

### 3.1 How can bias in hiring algorithms be measured?

Measuring bias in hiring algorithms is a complex process that requires an understanding of technological advancement and the various factors of production. One method of assessing bias is the calculation of input bias (Ib) and technological scale bias (TS)[7], which can determine whether the technological change is equal or biased towards a certain factor[8]. Additionally, total bias[10] and measurement error bias[8] can be calculated using statistical methods that depend on the type of nonresponse. To measure bias at the sentence and discourse levels respectively, two different association tests may be used[2]. Election-by-election estimates of partisan bias can also be used to measure bias, for example, to assess whether parties 'out-bias' each other[2]. Furthermore, a checklist for measuring race bias can be used[3], and any measure of bias should satisfy two criteria[4]. Finally, measurement modeling is a useful tool for understanding and uncovering implicit constructs in the language of bias measurement[5].

### 3.2 What are the different types of bias that need to be addressed?

There are a number of different types of bias that need to be addressed within the realm of Artificial Intelligence. For instance, the measure of input bias (Ib) is important when it comes to assessing technological advancement[7]. This measure of technological scale bias (TS,) is given by the equation[2,15], and is commonly used in epidemiology to detect bias due to measurement error[12]. Additionally, the total bias needs to be calculated for individual applications, which includes measurement error bias, response propensity and non-response bias[13]. Two different association tests have been designed to measure bias, one at sentence level (intra-sentence), and the other at discourse level (inter-sentence)[11]. Furthermore, election-by-election estimates of partisan bias can be calculated[2], and a checklist has been provided to measure the race bias[3]. The criteria for measuring bias should satisfy the condition that if a set remains unchanged, it should be assigned a value of zero[4]. Finally, a language of measurement modeling has been introduced to uncover the implicit constructs that such systems rely on[5]. Thus, it is essential to take these various types of bias into account when developing and assessing AHSs to reduce any potential systemic discrimination.

### 3.3 What metrics are used to evaluate the performance of a hiring algorithm?

In order to evaluate the performance of a hiring algorithm, various metrics are used. For instance, the measure of input bias (Ib) and the measure of technological scale bias (TS,) are two such metrics used to assess technological advancement. Additionally, the measure of bias due to measurement error[2] is commonly used in epidemiology. Furthermore, when nonresponse is encountered, total bias needs to be measured[9]. For instance, two different association tests can be designed to measure bias at the sentence and discourse levels[10], with election-by-election estimates of partisan bias used to assess methods for measuring bias and responsiveness[12]. Moreover, before considering the measurement of race bias[13], a two-criterion approach for measuring bias must be adopted[14]. Finally, to uncover the implicit constructs of a hiring algorithm, measurement modeling can help[15]. Thus, to assess the performance of a hiring algorithm, a variety of metrics are used to evaluate the fairness and equity of recruitment practices.

## 4. Model Design

### 4.1 Feature engineering

**Linguistic features**

Several linguistic features were used for each token. Each feature represents a characteristic property of the word. (**Table 1)** presents an explanation for each linguistic feature.

**Semantic features**

We utilized various state-of-the-art pre-trained word embeddings as textual features for the machine learning classifiers. The different word embeddings which were used are: Word2Vec (Mikolov et al. 2013), BERT (Devlin et al. 2019), ELMo (Peters et al. 2018), GloVe (Pennington et al. 2014), Flair (Akbik et al. 2018) and FastText (Bojanowski et al. 2017). Pre-trained word embeddings were used because the word embeddings trained on the EMSCAD dataset did not

demonstrate sufficient semantic quality due to the smaller size of the dataset. (**Table 2**) shows the pre-trained models used for each word embedding.

For each token, the word embedding vectors were extracted from the corresponding word embedding model using the FlairFootnote1 library.

**Feature selection**

The aforementioned linguistic features were combined with one of the six semantic features (word embedding) to produce a unique feature set. As a result, six unique feature sets were produced as input to the machine learning classifiers.

**4.2 Machine learning classifiers**

The machine learning classifiers were trained using the six unique feature sets on the training set of annotated job descriptions. The following classifiers were used:

- Support vector machine (SVM)
- Random Forest (RF)
- Logistic regression (LR)
- Decision tree (DT)
- Naive Bayes (NB)
- Multi-layer perceptron classifier (MLP)

For the baseline classifier, Scikit-learn's Dummy classifier was utilized. By performing parameter optimization using GridSearch, Footnote2 it was possible to search for the optimal parameters for all the machine learning classifiers. For all the classifiers, the maximum iterations were increased to infinity to ensure that the models are able to converge. All parameters, including the default parameters utilized for model training, are presented in (**Table 3)**.

**Table 1:** Linguistic features.

| Feature | Explanation |
|---|---|
| token.pos | Coarse-grained part-of-speech from the Universal POS tag set |
| token.ent.type | Named entity type |
| token.is.alpha | Does the token consist of alphabetic characters? |
| token.is.ascii | Does the token consist of ASCII characters? |
| token.is.digit | Does the token consist of digits? |
| token.is.lower | Is the token in lowercase? |
| token.is.upper | Is the token in uppercase? |
| token.is.title | Is the token in titlecase? |
| token.is.punct | Is the token punctuation? |
| token.is.space | Does the token consist of whitespace characters? |
| token.like.num | Does the token represent a number? |
| token.is.oov | Is the token out-of-vocabulary? |
| token.is.stop | Is the token part of a "stop list"? |
| token.lang | Language of the parent document's vocabulary |
| token.sentiment | A scalar value indicating the positivity or negativity of the token |
| token.len(word) | The length of the token |

**Table 2:** Word embeddings characteristics.

| Word embedding | Pre-trained model |
|---|---|
| Word2Vec | "en-glove" |
| BERT | "bert-base-cased" |
| ELMo | "medium" |
| GloVe | "en-glove" |
| Flair | "news-forward-fast" |
| FastText | "cc.en.300.bin" |

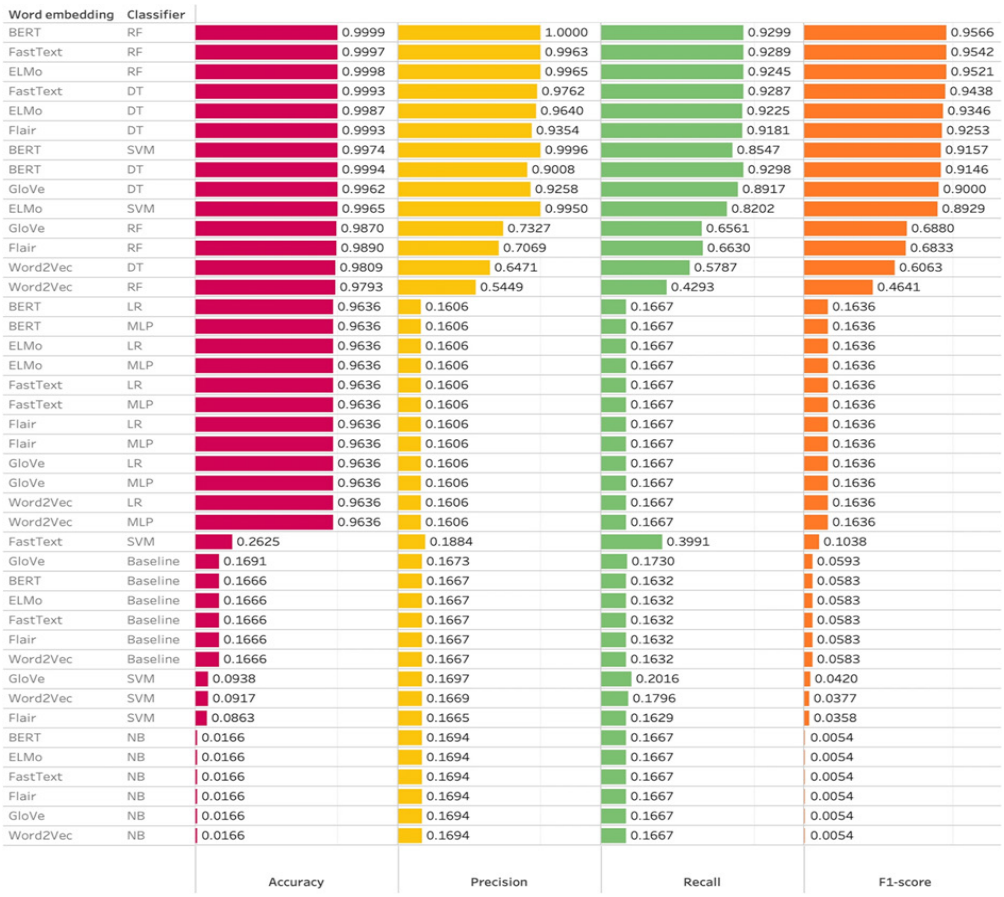
**Table 3:** Parameter grid.

| Baseline | SVM | LR | DT | RF | MLP |
|---|---|---|---|---|---|
| strategy = uniform | C = 10 | penalty = l2 | criterion = gini | n_estimators = 100 | hidden_layer_size = (100,) |
| constant = None | kernel = rbf | dual = False | splitter = best | criterion = gini | activation = relu |
| | degree = 3 | tol = 1e-4 | max_depth = None | max_depth = None | solver = adam |
| | gamma = 1 | C = 1.0 | min_samples_split = 2 | min_samples_split = 2 | alpha = 0.0001 |
| | coef0 = 0.0 | fit_intercept = True | min_samples_leaf = 1 | min_samprles_leaf = 1 | batch_size = auto |
| | schrinking = True | intercept_scaling =1 | min_weight_fraction_leaf = 0.0 | min_weight_faction_leaf = 0.0 | learning_rate = constant |
| | probability = False | class_weight = None | max_features = None | max_features = auto | learning_rate_init = 0.001 |
| | tol = 1e-3 | solver = newton_cg | max_leaf_nodes = None | max_leaf_nodes = None | power_t = 0.5 |
| | cache_size = 200 | max_iter = infinite | min_impurity_decrease = 0.0 | min_impurity_decrease = 0.0 | max_iter = infinite |
| | class_weight = None | multi_class = auto | class_weight = None | bootstrap = True | shuffle = True |
| | max_iter = infinite | warm_start = False | ccp_alpha = 0.0 | oob_score = False | tol = 1e-4 |
| | decision_function_shape = ovr | l1_ratio = None | | warm_start = False | warm_start = False |
| | | | | class_weight = None | momentum = 0.9 |
| | | | | ccp_alpha = 0.0 | Nesterovs_momentum = True |
| | | | | max_samples = None | early_stopping = False |
| | | | | | validation_fraction = 0.1 |
| | | | | | beta_1 = 0.9 |
| | | | | | beta_2 = 0.999 |
| | | | | | epsilon = 1e-8 |
| | | | | | n_iter_no_change = 10 |
| | | | | | max_fun = 15,000 |

## 5. Results and Analysis

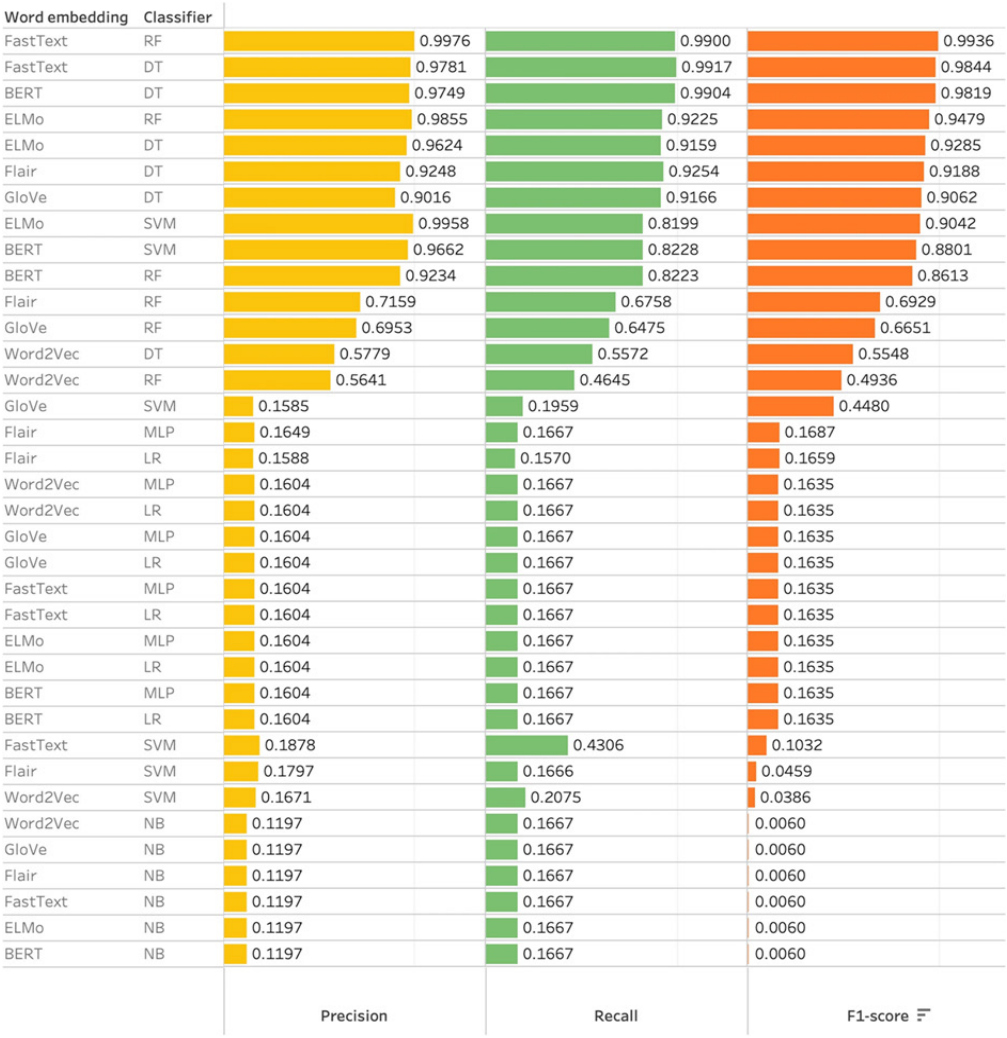
In this section, we present the results of various machine learning models on the EMSCAD dataset. The dataset was divided into 80% training and 20% testing set. The evaluation metrics: accuracy, precision, recall and F1-score were computed for each model. Figure 1 presents the evaluation metrics for various classifiers with different feature sets. The results indicate that the RF classifier with BERT word embeddings as textual feature achieved the best performance. This illustrates that contextual word embedding representations such as BERT had a superior performance over the non-contextual word embeddings such as FastText and Word2vec. We also observe that tree-based (Random Forest and Decision Tree) classifiers had a better performance in classifying biased and discriminatory language as compared to the remaining classifiers. Among the textual features, word embedding representations BERT, FastText and ELMo in combination with the RF classifier had the best performance. This was followed by FastText, ELMo and Flair word embeddings in combination with the DT classifier.

We further evaluate the various machine learning classifiers with different word embedding representations as features using tenfolds cross-validation. (**Figure 2)** presents the macro-averages of the precision, recall and F1-score over tenfolds cross-validation. The results of the tenfolds cross-validation indicate that the RF classifier with FastText word embeddings had the best performance. (**Figures 3, 4, 5 and 6)** present the individual results for accuracy, precision, recall and F1-score for the various models.



| Word embedding | Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| BERT | RF | 0.9999 | 1.0000 | 0.9299 | 0.9566 |
| FastText | RF | 0.9997 | 0.9963 | 0.9289 | 0.9542 |
| ELMo | RF | 0.9998 | 0.9965 | 0.9245 | 0.9521 |
| FastText | DT | 0.9993 | 0.9762 | 0.9287 | 0.9438 |
| ELMo | DT | 0.9987 | 0.9640 | 0.9225 | 0.9346 |
| Flair | DT | 0.9993 | 0.9354 | 0.9181 | 0.9253 |
| BERT | SVM | 0.9974 | 0.9996 | 0.8547 | 0.9157 |
| BERT | DT | 0.9994 | 0.9008 | 0.9298 | 0.9146 |
| GloVe | DT | 0.9962 | 0.9258 | 0.8917 | 0.9000 |
| ELMo | SVM | 0.9965 | 0.9950 | 0.8202 | 0.8929 |
| GloVe | RF | 0.9870 | 0.7327 | 0.6561 | 0.6880 |
| Flair | RF | 0.9890 | 0.7069 | 0.6630 | 0.6833 |
| Word2Vec | DT | 0.9809 | 0.6471 | 0.5787 | 0.6063 |
| Word2Vec | RF | 0.9793 | 0.5449 | 0.4293 | 0.4641 |
| BERT | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| BERT | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| ELMo | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| ELMo | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| FastText | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| FastText | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| Flair | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| Flair | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| GloVe | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| GloVe | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| Word2Vec | LR | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| Word2Vec | MLP | 0.9636 | 0.1606 | 0.1667 | 0.1636 |
| FastText | SVM | 0.2625 | 0.1884 | 0.3991 | 0.1038 |
| GloVe | Baseline | 0.1691 | 0.1673 | 0.1730 | 0.0593 |
| BERT | Baseline | 0.1666 | 0.1667 | 0.1632 | 0.0583 |
| ELMo | Baseline | 0.1666 | 0.1667 | 0.1632 | 0.0583 |
| FastText | Baseline | 0.1666 | 0.1667 | 0.1632 | 0.0583 |
| Flair | Baseline | 0.1666 | 0.1667 | 0.1632 | 0.0583 |
| Word2Vec | Baseline | 0.1666 | 0.1667 | 0.1632 | 0.0583 |
| GloVe | SVM | 0.0938 | 0.1697 | 0.2016 | 0.0420 |
| Word2Vec | SVM | 0.0917 | 0.1669 | 0.1796 | 0.0377 |
| Flair | SVM | 0.0863 | 0.1665 | 0.1629 | 0.0358 |
| BERT | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |
| ELMo | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |
| FastText | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |
| Flair | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |
| GloVe | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |
| Word2Vec | NB | 0.0166 | 0.1694 | 0.1667 | 0.0054 |

**Figure 1:** Machine learning models performance metrics.



| Word embedding | Classifier | Precision | Recall | F1-score |
|---|---|---|---|---|
| FastText | RF | 0.9976 | 0.9900 | 0.9936 |
| FastText | DT | 0.9781 | 0.9917 | 0.9844 |
| BERT | DT | 0.9749 | 0.9904 | 0.9819 |
| ELMo | RF | 0.9855 | 0.9225 | 0.9479 |
| ELMo | DT | 0.9624 | 0.9159 | 0.9285 |
| Flair | DT | 0.9248 | 0.9254 | 0.9188 |
| GloVe | DT | 0.9016 | 0.9166 | 0.9062 |
| ELMo | SVM | 0.9958 | 0.8199 | 0.9042 |
| BERT | SVM | 0.9662 | 0.8228 | 0.8801 |
| BERT | RF | 0.9234 | 0.8223 | 0.8613 |
| Flair | RF | 0.7159 | 0.6758 | 0.6929 |
| GloVe | RF | 0.6953 | 0.6475 | 0.6651 |
| Word2Vec | DT | 0.5779 | 0.5572 | 0.5548 |
| Word2Vec | RF | 0.5641 | 0.4645 | 0.4936 |
| GloVe | SVM | 0.1585 | 0.1959 | 0.4480 |
| Flair | MLP | 0.1649 | 0.1667 | 0.1687 |
| Flair | LR | 0.1588 | 0.1570 | 0.1659 |
| Word2Vec | MLP | 0.1604 | 0.1667 | 0.1635 |
| Word2Vec | LR | 0.1604 | 0.1667 | 0.1635 |
| GloVe | MLP | 0.1604 | 0.1667 | 0.1635 |
| GloVe | LR | 0.1604 | 0.1667 | 0.1635 |
| FastText | MLP | 0.1604 | 0.1667 | 0.1635 |
| FastText | LR | 0.1604 | 0.1667 | 0.1635 |
| ELMo | MLP | 0.1604 | 0.1667 | 0.1635 |
| ELMo | LR | 0.1604 | 0.1667 | 0.1635 |
| BERT | MLP | 0.1604 | 0.1667 | 0.1635 |
| BERT | LR | 0.1604 | 0.1667 | 0.1635 |
| FastText | SVM | 0.1878 | 0.4306 | 0.1032 |
| Flair | SVM | 0.1797 | 0.1666 | 0.0459 |
| Word2Vec | SVM | 0.1671 | 0.2075 | 0.0386 |
| Word2Vec | NB | 0.1197 | 0.1667 | 0.0060 |
| GloVe | NB | 0.1197 | 0.1667 | 0.0060 |
| Flair | NB | 0.1197 | 0.1667 | 0.0060 |
| FastText | NB | 0.1197 | 0.1667 | 0.0060 |
| ELMo | NB | 0.1197 | 0.1667 | 0.0060 |
| BERT | NB | 0.1197 | 0.1667 | 0.0060 |

**Figure 2:** Cross-validated Machine learning performance metrics.

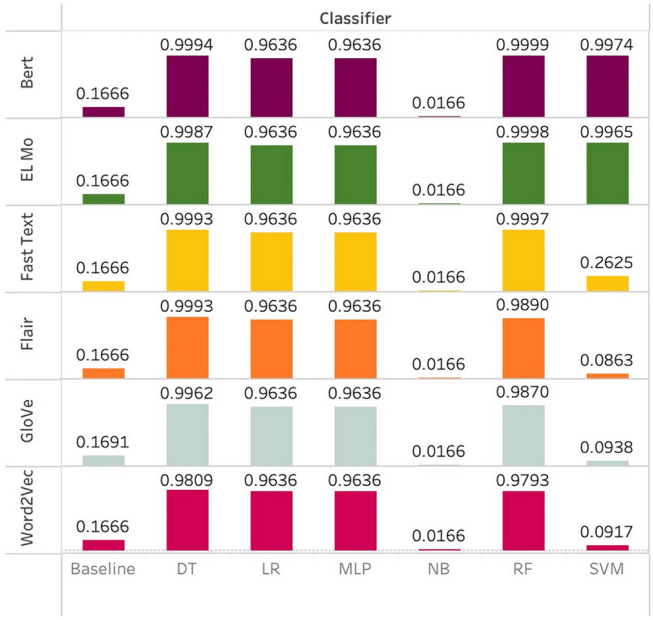



**Figure 3:** Machine learning models-accuracy.

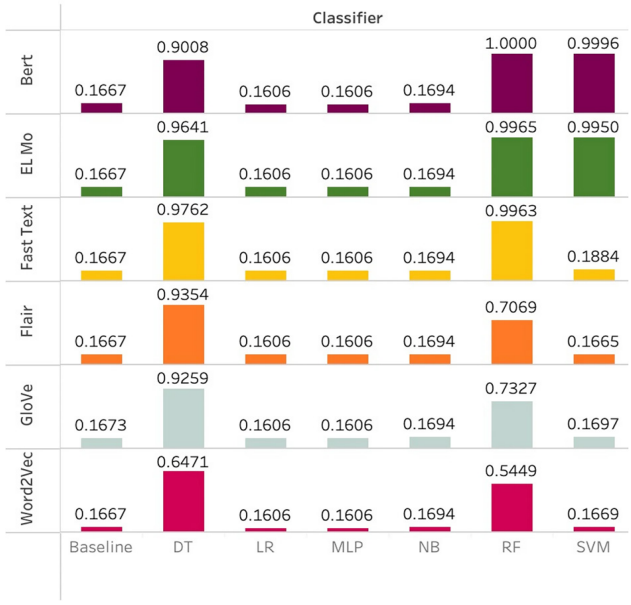



**Figure 4:** Machine learning models-precision.

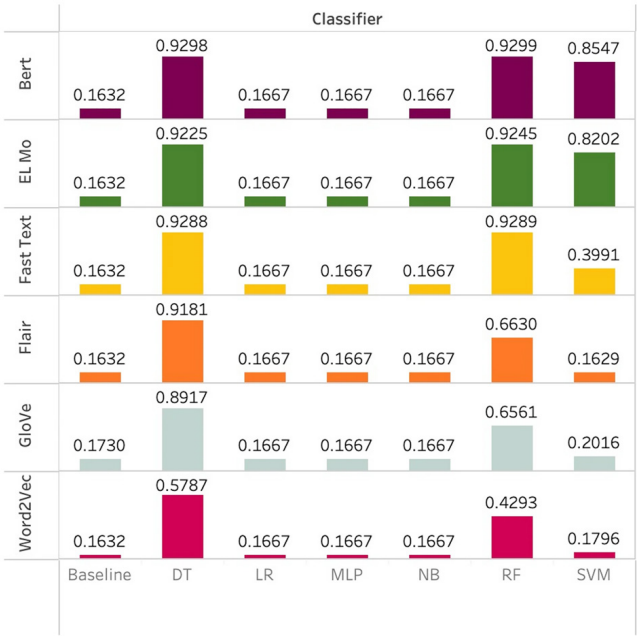



**Figure 5:** Machine learning models-recall.

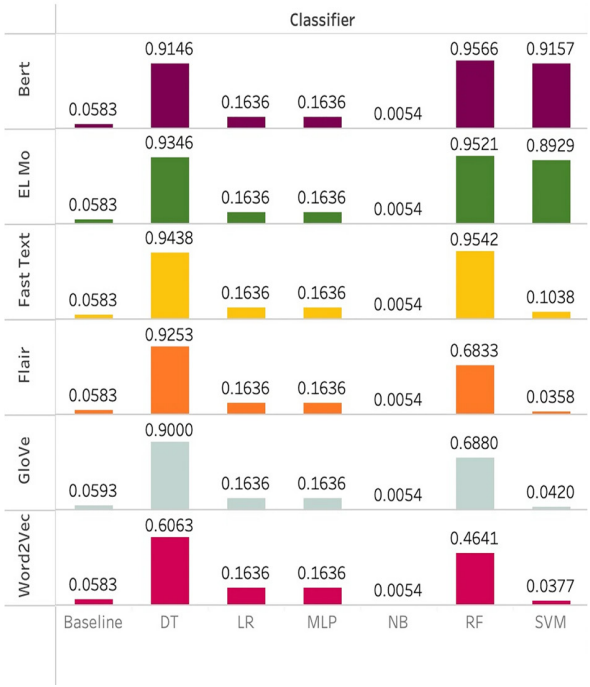



**Figure 6:** Machine learning models-F1.

Full size image **(Figures 7 and 8)** present the confusion matrices of the two best performing models: (1) RF classifier with FastText word embeddings and (2) RF classifier with BERT word embeddings. The results in **(Figs. 7 and 8)** indicate that all the five classes of biased and discriminatory language were distinguishable from each other.

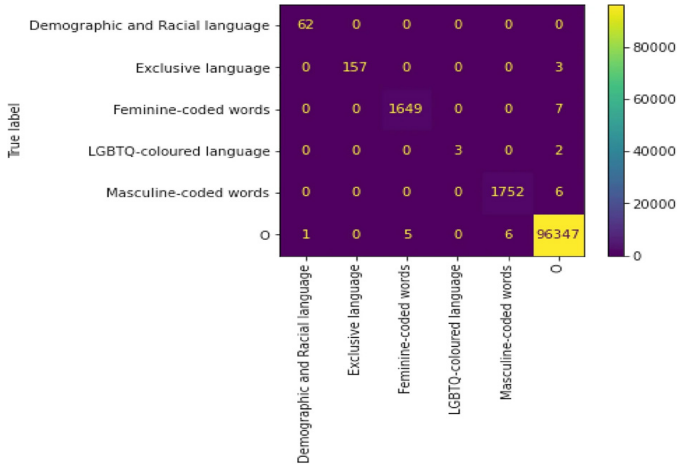



**Figure 7:** Confusion matrix Random Forest-FastText.

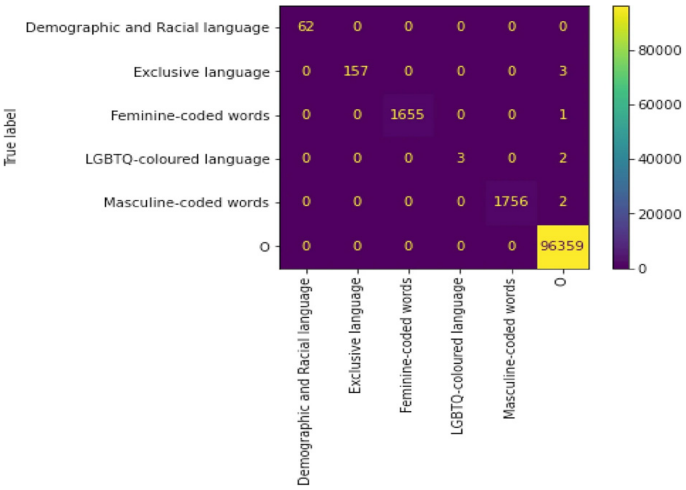



**Figure 8:** Confusion matrix Random Forest-BERT.

Full size image We observe a linear improvement in the performance of a sample of our models as training size increases in our primary experiment. We wanted to see if the behavior was by any chance related to these particular models or statistically grounded across all our implemented models. To validate this improvement, we included additional data (3000 additional job descriptions) and ran a new experiment with for the lightweight classifiers (DT, LR and NB).

The results obtained from the new experiment are shown in Fig. 9 for the regular models (80% training set and 20% test set), and Fig. 10 for the tenfold cross-validated models. We see that the DT classifier which uses BERT word embeddings produced the best performance: 0.98977, 0.99587 and 0.99277 for the precision, recall, and F1-score respectively. In fact, when compared to the result initially obtained by the best performing model in our first experiment, i.e., BERT—RF as shown in Fig. 1, we can see an improvement since the previous performance scores obtained were 0.98557, 0.98862 and 0.98544 equally for the precision, recall, and F1-score respectively.

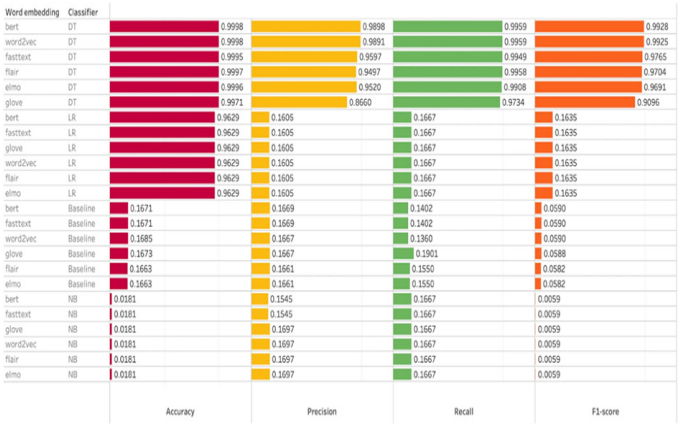



**Figure 9:** Extended machine learning models performance metrics.

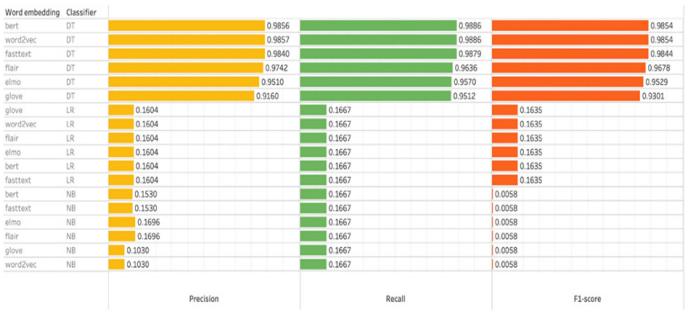



**Figure 10:** Extended cross-validated Machine learning performance metrics.

This analysis reinforces our belief that even the strong performance we have observed across the board can be further improved. However, it is not currently clear if, by any stroke of chance, the data samples utilized for evaluation might have been simplistic in the sense that they represent trivial cases. To the best of our knowledge, we have avoided cherry-picking by carrying out extensive random-sampling to select the evaluation set. However, given the small ratio of the entire evaluation set when compared to the size of our full data, we may not confidently rule out that this could have had an impact. In any case, we are leaving this for future study where we hope to perform a more comprehensive experiment on our entire dataset including carrying out an extensive ablation study with error analysis on the result.

## 7. Conclusion

By combining machine learning techniques, statistical analysis, and ethical considerations, our research methodology offers a holistic approach to bias mitigation in hiring algorithms. This comprehensive framework enables us to identify, measure, and mitigate both overt and subtle forms of bias, promoting equitable and fair candidate selection practices. The synergy of these methodologies ensures that bias is not only detected but also actively addressed to enhance diversity and inclusion in the recruitment process.

The integration of machine learning techniques into the research methodology plays a pivotal role in detecting, quantifying, and mitigating bias in hiring algorithms. By utilizing a combination of supervised learning, de-biasing strategies, and fairness-aware representation learning, this approach ensures that the hiring algorithm not only improves its overall performance but also adheres to ethical considerations and promotes fairness and inclusivity in candidate selection.

The incorporation of statistical analysis within the research methodology is instrumental in quantifying and understanding biases within hiring algorithms. By leveraging regression analysis, propensity score matching, A/B testing, and group-based disparities analysis, this approach offers a systematic and data-driven means to uncover both overt and subtle biases in candidate selection processes. These analyses contribute to a nuanced understanding of bias and provide insights into the effectiveness of bias mitigation strategies.

Ethical considerations play a pivotal role in shaping the research methodology for bias mitigation in hiring algorithms. By incorporating interdisciplinary perspectives, ensuring informed consent, promoting transparency, and fostering ongoing improvement, this approach aligns the research with ethical principles. Ultimately, the integration of ethical considerations helps create a framework that not only reduces bias but also upholds fairness, transparency, and social responsibility in the design and deployment of automated hiring systems.

The proposed framework for fostering diversity and inclusion in the workplace is necessary to ensure that all employees are treated fairly and equitably. An effective framework should consider both the technical and legal aspects of recruitment and hiring, including laws and regulations that prohibit discrimination[2]. It should also include measures to prevent discrimination and ensure that every individual is given equal opportunities[43]. This is the only way to ensure that organizations can create a workforce that is diverse in terms of gender, ethnicity, age, and other factors[44]. Furthermore, organizations can use the framework to ensure that the recruitment and hiring process is transparent and fair[5]. This would help them create a positive work environment and foster inclusion and collaboration among employees[46]. A successful framework should also be capable of providing support and guidance to employees during the recruitment and hiring process[7]. Ultimately, the proposed framework should be comprehensive enough to provide a sense of closure and completeness to the recruitment and hiring process[8]. This would make sure that the process is conducted in a fair manner and that all applicants have an equal chance of being selected[9]. A proper conclusion should also be drawn upon the completion of the recruitment and hiring process[5] so that the desired outcomes can be achieved[5].

This paper has presented a machine learning approach to identify five major categories of bias and discriminatory

language in job advertisements. We prepared a list of unique biased and discriminatory terms after examining the literature on behavioural works related to bias in recruitment. This list was used to semi-automatically generate an annotated corpus by the tagging the biased language terms (using a gazetteer-based approach) in the job advertisements of the publicly available Employment Scam Aegean Dataset, EMSCAD. This annotated corpus was used to train state-of-the-art machine learning classifiers to identify five different categories of biased and discriminatory language. We utilized a combination of linguistic features and most recent state-of-the-art word embedding representations as textual features to capture the natural language semantics of biased language. These features were fed into the machine learning classifiers. The results indicate that the Random Forest classifier with FastText word embeddings achieved the best performance with tenfold cross-validation. Overall, this work presents a major contribution in the attention phase of hiring and empowering recruiters by identifying and classifying discriminatory language in job advertisements using a machine learning-based approach. The output of this tool can be used to flag biased and discriminatory language and encourage recruiters to write more inclusive job advertisements.

The findings of this research paper highlight the critical issue of bias in hiring algorithms and the need for effective strategies to mitigate it. The study reveals that while automated hiring systems (AHSs) are being used to detect and address discrimination against protected groups, claims of 'bias mitigation' are rarely scrutinized and evaluated. The study emphasizes the importance of developing fair and equitable recruitment practices to promote greater diversity and inclusion in the workforce. The proposed framework for reducing bias in hiring algorithms aims to ensure that decision-making is done fairly and transparently, and provides support and guidance to employees during the recruitment and hiring process. However, the study also acknowledges that bias mitigation comes at a cost of efficiency and accuracy. The study suggests that ethical considerations should guide the development and implementation of AI-enabled recruitment practices, and that diverse groups should be represented in the development and testing of hiring algorithms. The study also highlights the importance of research methodology, including machine learning and statistical analysis, in providing greater accuracy and precision in results. Ultimately, this research contributes to the ongoing advancement of knowledge in the field of bias in hiring algorithms and provides valuable insights into effective strategies for mitigating bias and promoting equity in the workforce. Future research should continue to explore these issues and further develop the proposed framework for reducing bias in hiring algorithms. In conclusion, the recruitment and hiring process requires a proper conclusion to achieve the desired outcomes. This research paper highlights the critical issue of bias in hiring algorithms and the need for effective strategies to mitigate it. While automated hiring systems are being used to detect and address discrimination against protected groups, claims of bias mitigation are rarely evaluated. Fair and equitable recruitment practices are essential to promote greater diversity and inclusion in the workforce.

## 8. References


1. Monedero JS, Dencik L, Edwards L. What does it mean to 'solve' the problem of discrimination in hiring?: social, technical and legal perspectives from the UK on automated hiring systems. FAT 2020;458-468.
2. Hufthammer KT, Aasheim TH, Ånneland S, Brynjulfsen H, Slavkovik M. Bias mitigation with AIF360: A comparative study. NIK 2020;1:1-12.
3. Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: evaluating claims and practices. FAT 2020; 469-481.
4. Vasconcelos M, Cardonha C, Gonçalves B. Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions. AIES 2018;323-329.
5. Yarger L, Payton FC, Neupane B. Algorithmic equity in the hiring of underrepresented IT job candidates. Emerald Publishing Limited 2020;44(2):383-395.
6. Lyth AK. Challenging Biased Hiring Algorithms. Oxford Journal of Legal Studies 2021;41(4):899-928.
7. Stevenson R. Measuring technological bias. American Economic Association 1980;70(1):162-173.
8. Doraszelski U, Jaumandreu J. Measuring the bias of technological change. Journal of Political Economy 2018;126(3):1027-1084.
9. Olson K. Survey participation, nonresponse bias, measurement error bias, and total bias. Public Opinion Quarterly 2006;70(5):737-758.
10. Nadeem M, Bethke A, Reddy S. StereoSet: Measuring stereotypical bias in pretrained language models. Computation and Language 2020.
11. Jackman S. Measuring electoral bias: Australia, 1949-93. British Journal of Political Science 1994;24(3):319-357.
12. Cavazos JG, Phillips PJ, Castillo CD, O'Toole AJ. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. IEEE Trans Biom Behav Identity Sci 2021;3(1):101-111.
13. Grofman B. Measures of bias and proportionality in seats-votes relationships. Political Methodology 1983;9(3):295-327.
14. Jacobs AZ, Blodgett SL, Barocas S, Daumé H, Wallach H. The meaning and measurement of bias: lessons from natural language processing. FAT 2020: 706.
15. Edenhofer O, Knopf B, Barker T, et al. The economics of low stabilization: model comparison of mitigation strategies and costs. International Association for Energy Economics 2010;31(1):11-48.