Evaluating Discrimination Bias In AI Decision Making Systems For Personnel Selection

MSc Research Project
Data Analytics

Viktor Avgustin
Student ID: x20141432

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Viktor Avgustin
Student ID:	x20141432
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15/08/2022
Project Title:	Evaluating Discrimination Bias In AI Decision Making Systems For Personnel Selection
Word Count:	6372
Page Count:	23

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:	Viktor Avgustin
Date:	16th September 2022

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):	

Evaluating Discrimination Bias In AI Decision Making Systems For Personnel Selection

Viktor Avgustin
x20141432

Abstract

This research work aims at establishing a blueprint for conducting an ethical audit on Artifical Intelligence (AI) algorithms which are used for decision making when pre-screeing job applicant resumes to be considered for vacant positions. AI is widely used in industry for pre-screening CVs due to the cost-savings and ability to filter through large quantities of applicants. However, over reliance on these type of AI systems creates various ethical considerations. Most AI algorithms employed in CV screening are akin to "black boxes" in that they lack transparency in the way decisions are made. This research work highlights the importance of incorporating ethical considerations when building and training the algorithms in order to ensure that the production version is free of any discriminatory bias. In order to illustrate this, three classifier systems are built - kNN, Linear SVM and CNN to match CV of job applicants with job descriptions. The results are compared by adding gender as a sensitive variable to determine if any of the algorithms are bias towards gender e.g. selecting a larger proportion of a certain gender. The paper finds that there is a wide variation in the gender proportion across the three classifiers for the same job category. This indicates that a particular gender may be at a disadvantage based on the classifier used in the selection process. This paper advocates the building of robust classifiers which incorporate discriminatory variables in the training process to ensure that bias is eliminated when the classifier is deployed.

1 Introduction

Recruiting the most suitable staff for positions is critical for the existence of any organization. Automation of the hiring process is a natural evolution and certainly a fact for many organization of any size. Organizations employ the assistance of Artifical Intelligence (AI) and Machine Learning (ML) techniques to pre-screen resumes/Curriculum Vitae(CV) of prospective candidates. This allows organization to realize significant cost savings, in that a physical person is not paid a wage to scan a large volume of resumes (Gonzalez et al.; 2019) . In addition, the organizations are able to go through larger volumes of resumes and determine the most suitable candidate for a particular role. This is done by matching potential candidate CVs with a skill-set based on a job description criteria. However, this automation transfers a large portion of the responsibility on to the system and any flows with the algorithm can result in bias. This bias in turn can result in the business not acquiring the right talent for a position or even legal action from a persistent discriminatory pattern in the hiring process.

An AI hiring system may not perform as expected which can result in staff which does not possess the required skills for the position being progressed to the next round of interviews. This can be as a result of the AI/ML algorithm possessing a bias which has been incorporated during the training process. An example of a bias is when the system responds to particular key words, a candidate may be able to inundate their resume with potential keywords therefor fooling the system in believing that the candidate possesses those skills. The result is a system which is not fair to all candidates and the organization may be exposed to negative legal action because of its practices. In addition, more skilled candidate who have only listed keywords which are only applicable to their skill-set may be disregarded.

AI systems can be akin to "black boxes" making it nearly impossible to detect if a system is discriminating based on race, nationality, gender or against any minority group (Gonzalez et al.; 2019). There is evidence that discimination based on ethnicity and national origin is found in hiring practices (Oreopoulos; 2011). This may lead to potentially costly legal action against organizations, in addition to negative publicity and tainted corporate public image. The objective of this research paper is to gain insight into bias and discrimination incorporated in AI/ML decision making systems for personnel selection. Due to the large costs and benefits associated with hiring the right personnel for an organization, ensuring that the algorithm which sifts through the resume pool is performing as expected is critical to the company mission.

The AI/ML decision system will be subjected to an ethical audit to determine if it is fair and free of any bias. An AI decision making system is built which evaluates applicant CVs based on a skills criteria extracted from job profiles. The information from CVs is extracted by using Natural Language Process (NLP), which is then mapped into vectors and matched against a skill from a job description. There are three systems which are built by using Convolution Neural Networks (CNN), k-Nearest Neighbors (kNN) and Linear Support Vector Machine (SVM).

The research question being answered is the following:

Does CNN outperform kNN and Linear SVM in terms of reducing discrimination bias in employee skill based recommender systems?

Further to answering the research question, mitigants to reducing the discrimination bias of those classifiers are explored.

AI-ML decision systems are widely used in industry for personnel selection. Employing the CNN based decision making system significantly reduces bias and improves accuracy, which can contribute to organization being able to select the right staff.

Section 2 encompasses a review of the latest advances in decision making classifiers utilised for employee selection. It also summarizes literature which evaluates AI/ML decision making systems from an ethical perspective in an effort to eliminate bias and discrimination. Section 3 describes the research methodology employed. The architecture of the testing environment is explored in detail, the model employed and implementation is described in Section 4. Section 5 provides a summary and evaluation of the results. Section 6 is the conclusion of the paper and an exploration of possible future development of the topic.

2 Related Work

The first section of the survey examines systems focusing on employee behavior which operate in a similar fashion to the resume classifiers in that they extract skills and match them against requirements. The second section of the literature review focuses on the latest development in AI/ML recommendations systems used in decision making for resume selection. A limitation of the survey is that systems which are used commercially are proprietary and the details of how they operate are not publicly available. The third section focuses on Convolution Neural Networks (CNN) and its advances into text classification are examined in detail, along with feature selection and limitations. The last section of the literature review focuses on studies of fairness and eliminating bias in AI/ML algorithms and proposed methodologies of conducting an ethical bias.

2.1 A Recommender System for Employee Behaviour

Human Resources (HR) departments often build profiles for roles in an organization, which aids them in determining the right skill-set required from an applicant to be able to perform in a given role. Esmaeilzadeh et al. (2016) has created a set of profiles and employees hybrid clustering and optimisation models to evaluate the level of training for each employee and subsequently recommend whether any additional training is required. Fast intuitive clustering approach (FICA) and K-means are used to classify employees into the relevant categories of no training required, requires training and training necessary. While the design of the experiment lends itself to replication, the evaluation of the results is not clear and as a result it is unclear how well the proposed recommendation system performs. A similar design of grouping the skills extracted from CVs by using recommender systems is used in this research paper. An in-depth evaluation of the system is conducted along with an ethical audit to determine fairness and potential bias of the system.

Similar recommender systems are also employed in industry to determine the likelihood of an employee leaving the company. Some of the most successful classifiers to be utilized to predict employee attrition are k-Nearest Neighbors (kNN), artificial neural networks (ANN), decision trees and logistic regression (Yedida et al.; 2018). Based on accuracy the kNN achieves the highest rate of accuracy of 94.32 and it is used in this project as a classifier. Limitations of the study is that it does not distinguish between different skill-sets or career paths, as some career paths might be in higher demand, might have limited opportunities in an organization or might be in an industry which requires movement. In order, to mitigate this, key variables such as the employment title and the skill-set of the employee are incorporated into the model.

2.2 Latest Development in AI/ML Recommendation Systems

(Bafna et al.; 2019) employ Natural Language Processing (NLP) techniques which convert information from CVs/Resumes into vectors. Those vectors are constructed by using synset grouping and dimension reduction techniques on both the job descriptions containing the skills required and the descriptions on the applicant's CV. The study successfully employs clustering techniques to match these characteristics by matching based on the semantic meaning of each word, as applicants may use different words to describe the same concept. Bafna et al. (2019) demonstrates the flows of over-reliance on clustering by terms used in a resume as applicants may purposefully include terms/skills which they are not fluent in but only include them to "trick" the algorithm. This is of a particular concern when it comes to IT skills when including a particular programming language may allow an applicant to gain entry into the next level of the hiring process where that applicant is not necessarily proficient or even have experience with the language included. This may result in applicants which are being selected for skills which are far removed from their main field of expertise. However, the grouping employed in the study by incorporating semantic relativity between ensures that even if different words are used to describe the same skill, this is captured. Subsequently frequency is assigned to each term. The study highlights the benefits of employing a Term frequency–Inverse document frequency (TF-IDF) which is used in this body of research.

In Table 1 an example table is provided.

Table 1: Latest AI/ML Recommender System for Personnel Selection

Author, Year	**Description**	**Results**
(Bafna et al.; 2019)	Semantic grouping of terms	Precision: 0.94
	and assigning frequency score	Recall 0.90
(Roy et al.; 2020)	Linear SVM Classifier	Accuracy: 0.7853
(Lin et al.; 2016)	CNN/LSTM	Precision: 0.7
(Jiechieu and Tsopze; 2021)	CNN	Precision: 0.9134
(Ali et al.; 2022)	Linear SVM Classifier	Accuracy: 96
(Gopalakrishna and Vijayaraghavan; 2019)	kNN/dynamic	Accuracy: 91.2

Feature selection is critical to the design of the experiment and as such the selection of those features has a critical role in the design process. Manual feature selection is used in the selection of features such as gender, age, details and degree information, along with the frequency that a candidate has switched employment. Similar words are clustered by using Word2Vec, which is a type of neural network (Lin et al.; 2016). Semantic features are created based on the resulting clusters. Lin et al. (2016) have extracted 380 semantic features which can be excessive and lead to curve fitting. The study uses Long short-term memory (LSTM) and CNN in the classification stage. However, these methods did not perform well and had a low precision of .7 and the study did not publish any accuracy figures.

Term Frequency-Inverse Document Frequency (TF-ID) is widely used for feature extraction (Roy et al.; 2020) (Ali et al.; 2022), along with the Natural Language Toolkit (NTLK), which is used to clean the data, tokenize, stem and lemmatize the words from the CVs and job descriptions. (Roy et al.; 2020) have compared k-Nearest Neighbors (kNN), Random Forest, Multinomial Naive Bayes and Logistic Regression classifiers and Linear Support Vector classifiers (Linear SVM), with the Linear SVM achieving the highest accuracy of 78.53. However, these techniques are unable to measure the level of proficiency that a candidate posses in a given skill. These techniques lend themselves to the caveats described in previous studies where a candidate may exploit the system by over-utilizing a given keyword.

A Resume Classification System (RCS) as proposed by Ali et al. (2022) compares a number of different Support Vector Machines (SVM), such as Linear, SVC, SGD and NuSVC to compare against other classifiers - Naive Bayes classifiers, KNN and Logistic Regression. Linear SVM has achieved a very high accuracy rate of 96 percent. A robust evaluation of the results was provided by comparing results by using a Confusion Matrix, F-Score, Recall, Precision and Accuracy. However, a very small sample of nearly 900 CVs were used in the study and there was no evaluation of any ethical bias aspect of the classifiers.

Other studies have used a more ’dynamic’ approach which does not focus solely on training the data. Gopalakrishna and Vijayaraghavan (2019) have proposed Logistic Regression and k-Nearest Neighbors to classify suitable candidates within domains by using information on the CVs such as job descriptions, interests and experience. The dynamic aspect of this study comes from an ’ensemble learning-based voting classifier’ which changes after a number of iterations by refining and retraining the classifers. This classifier has achieved a 91.2 percent accuracy which is a significant improvement over the 84.2 percent achieved by the classifiers without voting.

2.3 Convolution Neural Networks for Natural Language

Convolutional Neural Networks (CNN) have only started being employed for text classification in recent years. However, Jacovi et al. (2018) have shown CNN to be performing well by using separate activation patterns for filters and global max-pooling techniques.

In literature CNN has only very recently started being used for resume classification, to match a set of CVs against a list of skills extracted from job descriptions. Jiechieu and Tsopze (2021) have extracted skills from CVs from a sample of nearly 20k applicant CVs which were matched against a number of skills. The model has achieved a 91.34 percent precision. The neural classifier performs document encoding by using filters of size 1, 2 and 3. The clustering process is made more efficient by training of the classifier which improves the encoding, and in turn the resulting clusters are a better representation of the input document. The filter 1 is used which discovers most patterns or related words. Text features of size larger than 1 may not be identified, even if patterns do exist. This technique suffers from similar defects as the ones described before as they do not consider the context of the terms being used. In addition, the system was only applied to the IT sector, where specific languages or computer skills are the key words. Expanding this model to other industries may be beneficial to the body of knowledge.

CNN has also been used by Mridha et al. (2021) for ranking resumes that match a certain job specification. The study uses key words and again suffers from the same problem as it can be manipulated when keywords are included on the CV. The model has achieved a 74 percent accuracy but it has not been evaluated for any bias, fairness or any ethical concerns, which is the main objective of this research.

Attempts to overcome the problem of context have been made by including word embedding techniques. (Wings et al.; 2021) measures the impact that the context has on the classification model for skills extraction by employing shallow classifiers and linear machine learning algorithms. The study performs an in-depth comparison to other techniques used for word embedding and creates a ’context aware system for skill extraction’.

Gaur et al. (2021) have extracted education qualifications from CVs by using a semi-supervised approach. The proposed model used 3 layers starting with the word embedding, CNN, and a Bi-LSTM layer. However, the experiment has a very heavy reliance on manual annotation of the data and the authors have shown that manually annotating the data inputs imrpoves the results.

2.4 Defining bias in AI/ML Recommender Systems

The European Union's Directive 2002/73/EC, Article 21 of the Charter of Fundamental Rights and Protocol 12/Article 14 of the European Convention on Human Rights defines discrimination and anti-discrimination. The EU's framework on equal treatment and anti-discrimination define the ground for discrimination on the basis of sexual orientation, gender, nationality, ethnic origin and race, disability, age and religion or belief. The research examines bias based on gender and nationality/ethnic origin as only these characteristics are readily available in a CV.

Large volumes of candidates are filtered through a resume selection system and an embedded bias in the AI system which may have been overlooked while training the algorithm, may result in the system discriminating against minority groups. Žliobaitė (2017) proposes a framework to measure the discrimination bias and proposes a solution to reduce it. The mean difference method and the normalized difference methods are recommended to measure bias in cases of well balanced data between the different cases.

Resume selection AI systems may have embedded bias and discrimination built in when the algorithm is being trained and built. When the training dataset does not follow the same distribution as the population on which the AI system is being deployed, this leads to bias (Calders and Žliobaitė; 2013). Similarly, if there are errors in the data and the algorithm is unable to capture all the possible variations. Calders and Žliobaitė (2013) have also found that a differing economic cycle may also lead to bias, e.g. if the algorithm is trained in a recession cycle versus being deployed in an economic boom may cause bias when determining credit worthiness of candidates. As economic conditions change and skill-sets evolve, the system may not be able to rapidly adapt to the changing environment.

In resume selection, education is critical and AI systems may develop a bias towards educational institutions from a particular country, due to dataset being from a particular country. Sampling bias can be created if the training dataset is not representative of the population. This can also happen with modules in a university degree, where foreign modules are labeled differently and are therefor not being recognized by the AI, which leads to labeling bias.

2.5 Ethics and Ethical Audits of AI/ML Recommender Systems

AI systems are very commonly referred to as "black boxes", because they lack transparency in the way that they make decisions. Transparency is critical for measuring fairness of an AI algorithm. However, the definition of fairness and bias differs widely across the literature. Some large technology companies - such as Google or Meta have large departments which focus on the interpretation of fairness and determining bias (Landers and Behrend; 2022). Landers and Behrend (2022) proposes splitting the AI system into its various components and asking questions at each step with a view of detecting and eliminating bias. The AI system is split into its various components - the input data layer, the model design, development features and processes, all the way through to the output.

Robert et al. (2020) describes "fairness" and the study split fairness into the following types - distributive, procedural and interactional fairness. Distributive fairness is of particular interest in this project as it relates to equity and ensuring that equal inputs are matched with equal outputs. Procedural fairness focuses on the specific process used when reaching an outcome. In the case of AI systems, all the candidates need to go through the AI system before they an reach the next round. Leventhal et al. (1980) define the characteristics of procedural fairness as the following: ”consistency, unbiased suppression, represenativeness, correctability, accuracy and ethicality. Finally, interactional fairness is defined as the treatment received by the individual from the organization administering the AI system.

The lack of transparency and fairness of the AI system can lead to lack of trust in the system by the applicants. Gonzalez et al. (2019) finds that applicants are less favourably pre-disposed towards an AI/ML system which evaluates job applications as opposed to a human decision maker. This leads to negative view of the brand of the organization and lead to negative publicity. This further strenghtens the case for ensuring that the system is transparent and fair by performing an ethical audit.

3 Methodology

This section describes the methodology used in performing the experiments in this research work. The scientific process in this research paper follows the steps described. A pre-labeled dataset is used to train the models to match resumes to CVs. The second step is to collect resume and job description data, pre-process and clean it to create the datasets for evaluating the models. Gender information is extracted from the datasets and the classifier models are ran on the collected data. Subsequently, experiments are build to match resumes and job descriptions. Initially by using the words and subsequently more complex classification models which employ word2doc and TF-IDF feature extractions. The flow from data collection, processing, feature extraction and modelling is explained in Figure 1.

3.1 System Specification and Software

The research was conducted on an Processor Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, 2701 Mhz, 2 Core(s), 4 Logical Processor(s) with 16GB of Physical Memory and 32GB of virtual memory. The data collection and coding was performed in Python by utilising the PyCharm software and Google colab.

3.2 Data Collection

The dataset was built by scraping publicly available resume data from postjobfree.com and linkedin.com. Two datasets were scraped, one which contained many different job titles and another which was only targeted for specific job titles. Creating a custom dataset from a different source, allows the research to evaluate bias of the pre-built machine learning model (Gianfrancesco et al.; 2018). A publicly available labelled dataset is used to train the models.

3.2.1 Pre-Labelled Dataset

The pre-labelled dataset is collected from Kaggle. This dataset contains resumes and job title labels. It is used for training the models.

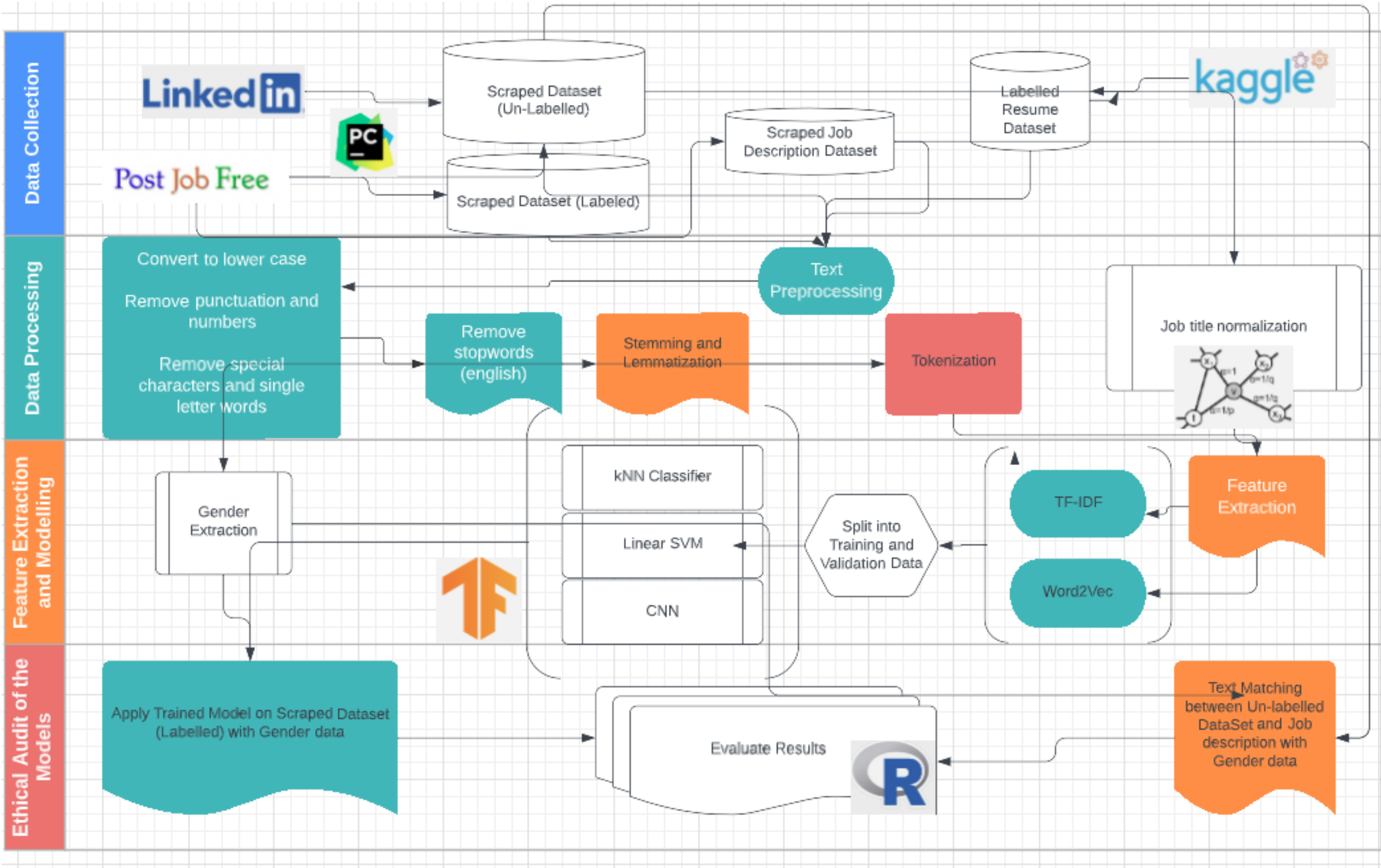

Data Collection
Data Processing
Feature Extraction and Modelling
Ethical Audit of the Models
LinkedIn
Post Job Free
Scraped Dataset (Un-Labelled)
Scraped Dataset (Labeled)
Scraped Job Description Dataset
Labelled Resume Dataset
kaggle
Convert to lower case
Remove punctuation and numbers
Remove special characters and single letter words
Text Preprocessing
Remove stopwords (english)
Stemming and Lemmatization
Tokenization
Job title normalization
Gender Extraction
kNN Classifier
Linear SVM
CNN
Split into Training and Validation Data
TF-IDF
Word2Vec
Feature Extraction
Apply Trained Model on Scraped Dataset (Labelled) with Gender data
Evaluate Results
Text Matching between Un-labelled DataSet and Job description with Gender data

Figure 1: Methodology Flow Chart

3.2.2 Collected Dataset (Un-labelled)

The datasets utilized in this project have been scraped from the internet by using linkedin.com and postjobfree.com. Scraping large volumes of resumes is particularly challenging as many websites such as linkedin.com and indeed.com, significantly limit the number of CV/Resumes that can be accessed on a daily basis. Several techniques have been employed to circumvent these limitations.

The scrapy library, along with the selenium library were employed for scraping. The Docker platfrom was used to run selenium and a linkedin crawler for the linkedin crawling. Techniques involved in the scraping of the data were the following:
1. Scraping from postjobfree.com by the use of the scrapy library. The technique employed was a timeout of 10 seconds between searches. A search was performed on a keyword profession e.g. ”Accountant”. Five search result pages were searched and resumes scraped from those. After a period another search was performed on a different keyword.
2. Searching on linkedin.com has many restrictions as the website allows to only search a little over 100 accounts per day, which are not in a person's contact list. However, there is no limitation on searches by name. In order, to generate linkedin names, a google.com search is performed for a given profession, similar to the postjobfree.com search. The search limits to only linkedin results by the following string: site:linkedin.com/in/. By using the BeautifulSoup library the names of linkedin profiles are extracted for the searches. Google allowed circa ten pages results to be searched at time, after which a captcha test

was manually passed by the researcher. Linkedin urls were also collected and those were used to perform a random scrapy crawl on linkedin.com.

The gathered list of names are input into a text file and a search by name is performed on linkedin.com by using the scrapy library. A few hours of timeouts were required between searches and the number of names searches were limited to 100. Different linkedin accounts were required to avoid the accounts being blocked by the system. The data scraped contains the name of the person, education data, employment data, along with the name and description of the particular role. Dates of employment and education were also gathered.

3.2.3 Collected Dataset (Labelled)

To allow the models to be tested on the collected datasets, a dataset is created by employing a similar technique to the collection of the un-labelled dataset. However, in the collection of the labeled dataset the search is fine-tuned to ensure that only labels from a specified list of labels is collected. This allows for the accuracy of the models to be tested on a diffierent set of data and also allows for any discrimination bias to be evaluated by adding the gender (Jiechieu and Tsopze; 2021).

3.3 Text Pre-processing and Dataset Preparation

The pre-processing and dataset preparation process consisted of converting to lower case, removing unnecessary characters, removing stop words, removing urls, also lemmatizing and tokenizing the words.

The Un-labelled dataset required some additional normalization of the job titles. Ensuring that generic labels are used e.g. 'senior accountant' is the same as 'accountant' has proven to be particularly challenging. Algorithms along with Node2Vec has been employed to reduce the number of job titles and ensure they fit into more generic categories.

3.4 Datasets

Three datasets are used in this research - a labelled datataset, a scraped unlabelled dataset and a scraped labeled dataset.

3.4.1 Pre-labelled Dataset

The pre-labelled dataset was downloaded from Kaggle[1]. It contains 962 resumes which are split into 25 labeled categories. Figure 2 represents the split of the dataset into the separate label categories.

[1] https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset

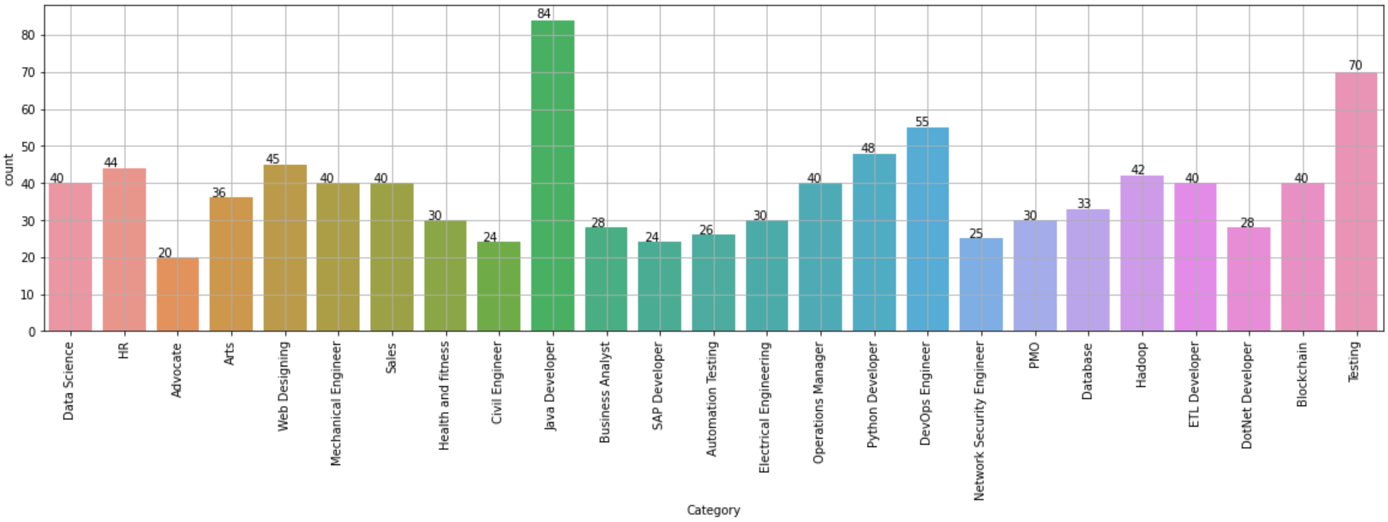
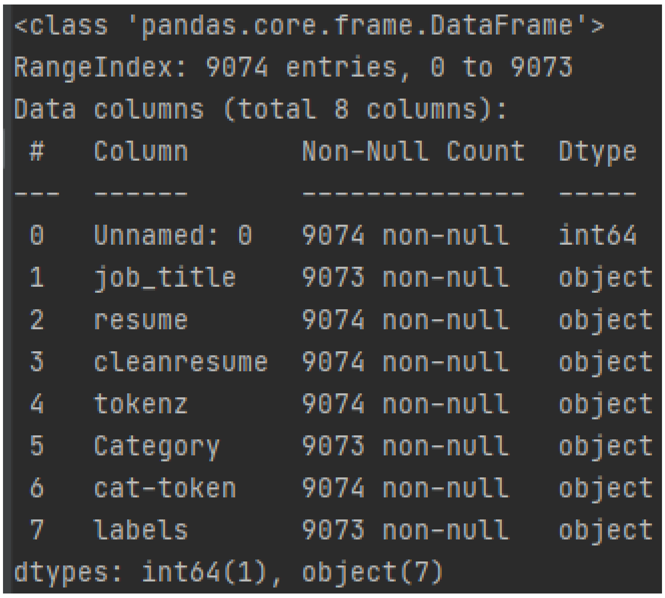


Figure 2: Kaggle labeled dataset

3.4.2 Scraped Dataset (Un-labelled)

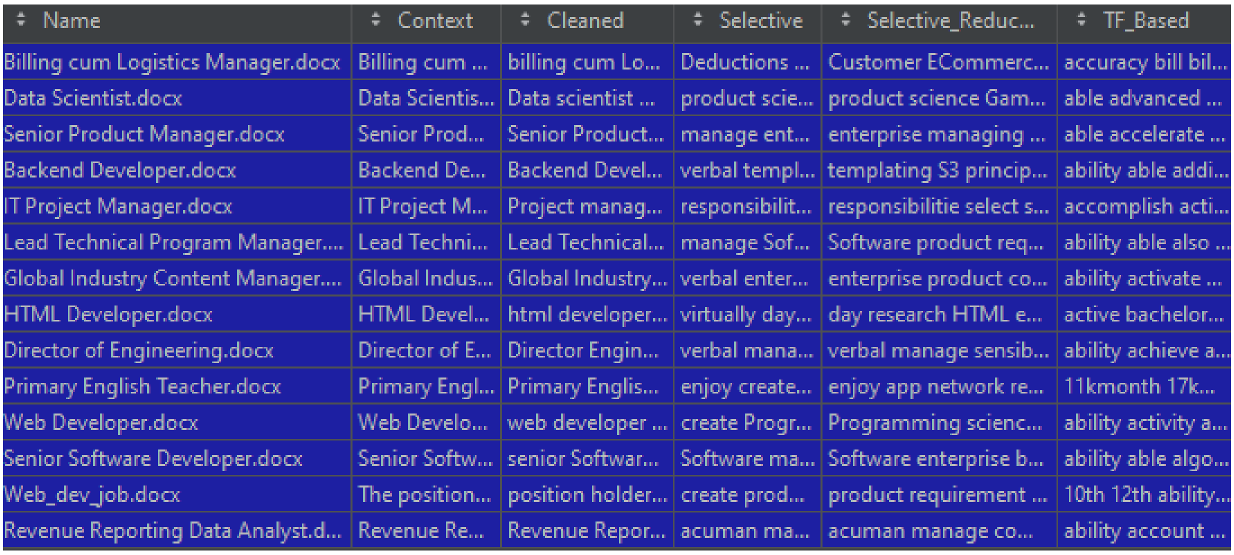
The created resumes dataset consists of 9074 resumes. The resume dataset is split into 8 columns. As shown in Figure 3.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9074 entries, 0 to 9073
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Unnamed: 0   9074 non-null   int64 
 1   job_title    9073 non-null   object
 2   resume       9074 non-null   object
 3   cleanresume  9074 non-null   object
 4   tokenz       9074 non-null   object
 5   Category     9073 non-null   object
 6   cat-token    9074 non-null   object
 7   labels       9073 non-null   object
dtypes: int64(1), object(7)
```

Figure 3: Resume Dataset description

The job title dataset consists of 14 job desciptions shown in Figure 4.



Name	Context	Cleaned	Selective	Selective_Reduc...	TF_Based
Billing cum Logistics Manager.docx	Billing cum ...	billing cum Lo...	Deductions ...	Customer ECommerc...	accuracy bill bil...
Data Scientist.docx	Data Scientis...	Data scientist ...	product scie...	product science Gam...	able advanced ...
Senior Product Manager.docx	Senior Prod...	Senior Product...	manage ent...	enterprise managing ...	able accelerate ...
Backend Developer.docx	Backend De...	Backend Devel...	verbal templ...	templating S3 princip...	ability able addi...
IT Project Manager.docx	IT Project M...	Project manag...	responsibilit...	responsibilitie select s...	accomplish acti...
Lead Technical Program Manager....	Lead Techni...	Lead Technical...	manage Sof...	Software product req...	ability able also ...
Global Industry Content Manager....	Global Indus...	Global Industry...	verbal enter...	enterprise product co...	ability activate ...
HTML Developer.docx	HTML Devel...	html developer...	virtually day...	day research HTML e...	active bachelor...
Director of Engineering.docx	Director of E...	Director Engin...	verbal mana...	verbal manage sensib...	ability achieve a...
Primary English Teacher.docx	Primary Engl...	Primary Englis...	enjoy create...	enjoy app network re...	11kmonth 17k...
Web Developer.docx	Web Develo...	web developer ...	create Progr...	Programming scienc...	ability activity a...
Senior Software Developer.docx	Senior Softw...	senior Softwar...	Software ma...	Software enterprise b...	ability able algo...
Web_dev_job.docx	The position...	position holder...	create prod...	product requirement ...	10th 12th ability...
Revenue Reporting Data Analyst.d...	Revenue Re...	Revenue Repor...	acuman ma...	acuman manage co...	ability account ...

Figure 4: Job Dataset description

3.4.3 Scraped Dataset (Labelled)

After preprocessing (cleaning, removing incorrect labels, removing resumes with unknown gender) the scraped labelled dataset contains 1856 resumes which are split into the same 25 categories as the Kaggle dataset. Figure 5 is a graphical depiction of the distribution of those categories.

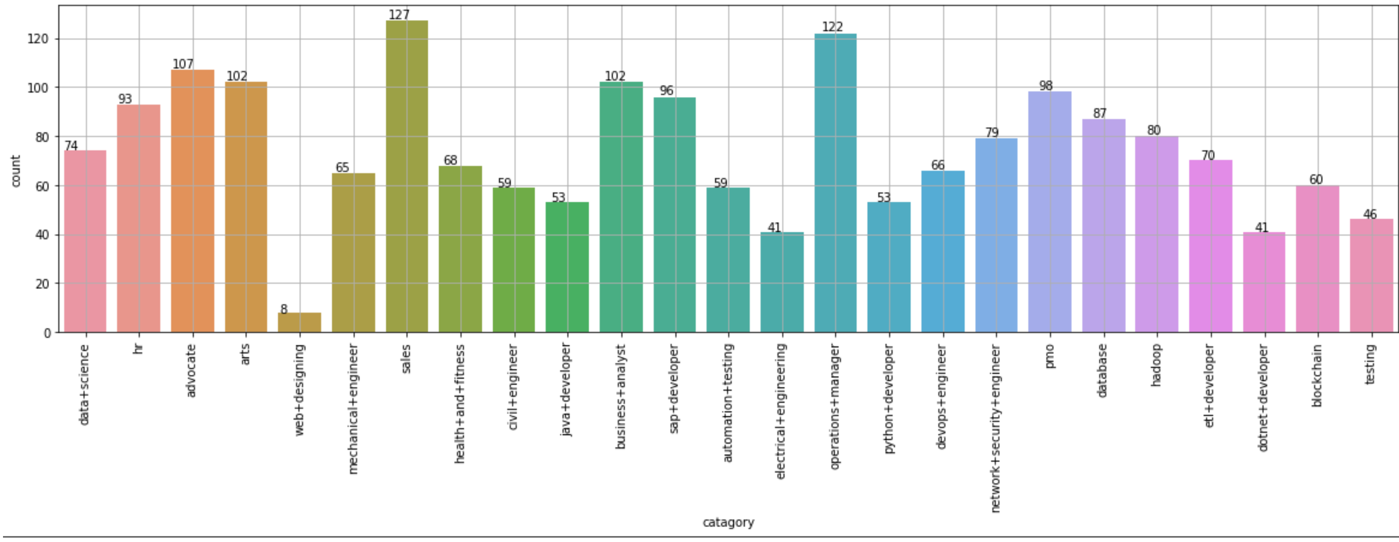


Figure 5: Scraped Labelled Dataset

3.5 Feature extraction

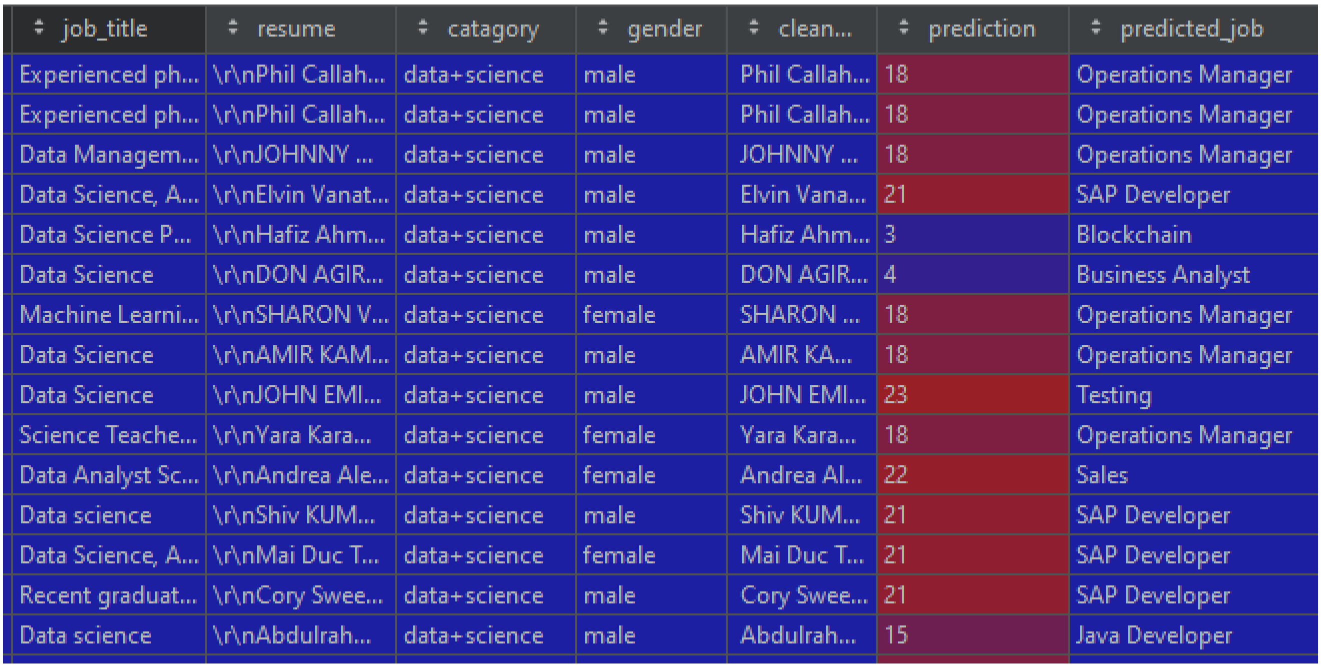
The NLTK library is used for text processing (Pan et al.; 2019). This methodology is a simplified version of the methodology used by Jiechieu and Tsopze (2021), which extracts skills from resumes and job descriptions. Word embedding are employed such as Word2Vec and ord frequency and inverse document frequency (TF-IDF) (Wang and Shi; 2022), used to create word vectors. These techniques are rather successful with word2vec being more complex as it is able to compute probability given the context of the word (Wang and Shi; 2022). Using n-grams for text classification is a rather common technique employed widely in the literature and that is why it is chosen as the preferred method in this body of research (Roy et al.; 2020).

3.6 Training the models

Once the features vectors are extracted from resumes, they are matched against a job description (job label) by using classifiers such as Linear SVM Classifier (Roy et al.; 2020), kNN (Gopalakrishna and Vijayaraghavan; 2019) and CNN (Jiechieu and Tsopze; 2021). These classifiers are initially trained on the Kaggle labeled dataset and subsequently deployed on the scraped labeled dataset, which also contains gender information. This is done in order to simulate a real world environment where a CV filtering algorithm is deployed into production to filter CVs for job interviews. In that case, the algorithm will be deployed to data that has not been used for training.

3.7 Match Resumes and Job Specifications

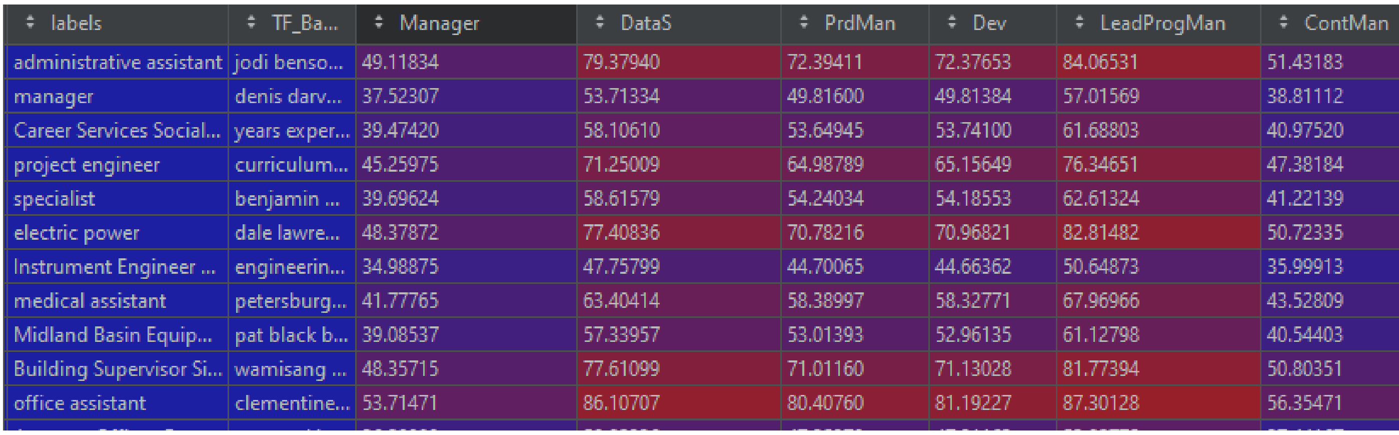
The classifiers are used to assign a label to a resume which identifies it as being selected for a particular position. This technique is being employed on the scraped labeled dataset.Figure 6 demonstrates the labelled dataset with the predicted job.



job_title	resume	catagory	gender	clean...	prediction	predicted_job
Experienced ph...	\r\nPhil Callah...	data+science	male	Phil Callah...	18	Operations Manager
Experienced ph...	\r\nPhil Callah...	data+science	male	Phil Callah...	18	Operations Manager
Data Managem...	\r\nJOHNNY ...	data+science	male	JOHNNY ...	18	Operations Manager
Data Science, A...	\r\nElvin Vanat...	data+science	male	Elvin Vana...	21	SAP Developer
Data Science P...	\r\nHafiz Ahm...	data+science	male	Hafiz Ahm...	3	Blockchain
Data Science	\r\nDON AGIR...	data+science	male	DON AGIR...	4	Business Analyst
Machine Learni...	\r\nSHARON V...	data+science	female	SHARON ...	18	Operations Manager
Data Science	\r\nAMIR KAM...	data+science	male	AMIR KA...	18	Operations Manager
Data Science	\r\nJOHN EMI...	data+science	male	JOHN EMI...	23	Testing
Science Teache...	\r\nYara Kara...	data+science	female	Yara Kara...	18	Operations Manager
Data Analyst Sc...	\r\nAndrea Ale...	data+science	female	Andrea Al...	22	Sales
Data science	\r\nShiv KUM...	data+science	male	Shiv KUM...	21	SAP Developer
Data Science, A...	\r\nMai Duc T...	data+science	female	Mai Duc T...	21	SAP Developer
Recent graduat...	\r\nCory Swee...	data+science	male	Cory Swee...	21	SAP Developer
Data science	\r\nAbdulrah...	data+science	male	Abdulrah...	15	Java Developer

Figure 6: Predicted Job on Labeled Dataset

A different technique is being applied to the scraped unlabelled dataset. The resumes are matched against job descriptions by using the text distance and a score is assigned to for each resume in a given job description (Kadhim; 2019). Figure 7 portrays a number of resumes (rows) and the assigned score for each job description (column).



labels	TF_Ba...	Manager	DataS	PrdMan	Dev	LeadProgMan	ContMan
administrative assistant	jodi benso...	49.11834	79.37940	72.39411	72.37653	84.06531	51.43183
manager	denis darv...	37.52307	53.71334	49.81600	49.81384	57.01569	38.81112
Career Services Social...	years exper...	39.47420	58.10610	53.64945	53.74100	61.68803	40.97520
project engineer	curriculum...	45.25975	71.25009	64.98789	65.15649	76.34651	47.38184
specialist	benjamin ...	39.69624	58.61579	54.24034	54.18553	62.61324	41.22139
electric power	dale lawre...	48.37872	77.40836	70.78216	70.96821	82.81482	50.72335
Instrument Engineer ...	engineerin...	34.98875	47.75799	44.70065	44.66362	50.64873	35.99913
medical assistant	petersburg...	41.77765	63.40414	58.38997	58.32771	67.96966	43.52809
Midland Basin Equip...	pat black b...	39.08537	57.33957	53.01393	52.96135	61.12798	40.54403
Building Supervisor Si...	wamisang ...	48.35715	77.61099	71.01160	71.13028	81.77394	50.80351
office assistant	clementine...	53.71471	86.10707	80.40760	81.19227	87.30128	56.35471

Figure 7: Resume and Job Description Scoring

3.8 Ethical audit

Gender is extracted from each CV by evaluating the first name of the applicant. In order to determine if there is any underlying bias by the classifier and evaluation of the results is performed. The proportion of male vs female applicants is evaluated for each score percentile to determine if a certain gender scores statistically significantly lower or higher for given job specifications.

The proposed methodology for determining the gender bias is the following as employed by (Alelyani; 2021), which suggests training the model on a dataset, predicting the class labels for each data point, applying the alternative function on the gender attribute, training the model on the alternative dataset and predicting the alternative predict label. The evaluation is completed by measuring the distribution of gender labels across the predicted and expected job labels.

4 Design Specification

The research paper aims at matching resumes of prospective candidates with job specifications by using a classifier. Once the classifications is performed and a score is determined for each resume for a given job description, the model is evaluated for ethical bias by adding the gender dimension. This is performed by calculating the number of male/female proportion for each job description based on the score they have received split into percentiles.

4.1 Feature Extraction

Features are extracted from the resumes by employing the TF-IDF algorithm. TF-IDF for a document is a score which is calculated by multiplying the term frequency of a word in a document by the inverse document frequency of the word across a set of documents. A very common word will have an inverse frequency very close to zero, where a rarely used word will have a inverse frequency of close to 1. A high score on TF-IDF indicates a more relevant word (Trstenjak et al.; 2014).

Another technique for word embedding is Word2Vec. It is also used to map words into vectors, however the word2vec vectors more closely resemble neural networks. The advantage of word2vec is that it is able to make inference on the meaning of a word in a text based on the occurrence in a corpus of text. Word2Vec allows for the option of building a cotinuous skip-gram model which is simple neural network with a hidden layer which predicts the probability of a given word from an input (Mikolov et al.; 2013).

Word2vec does not appear to be sensitive to the size of the embedding size as the papers examined have used arbitrary dimensions without affecting performance (Ilić et al.; 2018). The dimension size chosen is 100, as suggested in the literature that a smaller dimension is to be used for a classification task such as the one employed in this paper.

Node2Vec which is based on Word2Vec is used to normalize the job-titles in the un-labelled scraped dataset. This is done in an effort to reduce the number of job labels on that dataset.

GloVe - global vectors for word representation is used in the training of the CNN model. Glove is an unsupervised learning algorithm which used for vector representation of words.

Future reseach can benefit from optimisation of the parameters employed in the TF-IDF and Word2Vec vectorizers.

4.2 Gender Extraction

The gender is an important part in determining bias in the algorithm based on gender. The gender is extracted from both the labelled and un-labelled scraped datasets by employing the gender-guesser library. Gender is determined based on the first name of the applicant. Resumes where the gender can not be determine wit certainty are erased from the datasets.

4.3 Models

Three models are trained and applied to the collected labelled dataset - kNN, linear SVM and CNN.

4.3.1 K-Nearest Neighbors Algorithm (kNN)

The k-nearest neighbors (KNN) algorithm is used to build a model which is trained on the labeled Kaggle dataset. It is a supervised machine learning algorithm which is widely used for classification problems (Trstenjak et al.; 2014). The classification problem in this research work is the classification of scraped resumes into job titles.

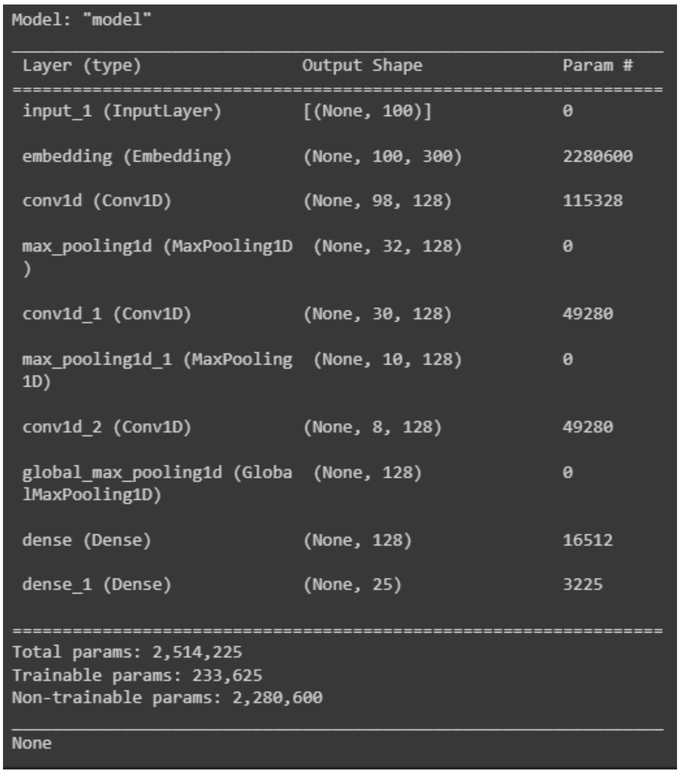
4.3.2 Linear Support Vector Machine (SVM)

Linear Support Vector Machine is a linear model which uses uses regression for classification problems. It has been used to solve classification problems for resume and job description matching (Roy et al.; 2020). Linear SVM creates a hyperplane which separates resumes into job titles.

4.3.3 Convolutional Neural Network (CNN)

This research paper employs a CNN to classify resumes into the correct job listing category. The design of the CNN model is shown in Figure 8 (Wang; 2018). The model employs GloVe for vectorization. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The Glove.840B.300d is used, which contains 840B tokens, 2.2M vocab and 300d vectors with a size of 2.03GB [2].

Training is performed by comparing the output of the model to the output label.



```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 100)]             0

 embedding (Embedding)       (None, 100, 300)          2280600

 conv1d (Conv1D)             (None, 98, 128)           115328

 max_pooling1d (MaxPooling1D  (None, 32, 128)          0
 )

 conv1d_1 (Conv1D)           (None, 30, 128)           49280

 max_pooling1d_1 (MaxPooling  (None, 10, 128)          0
 1D)

 conv1d_2 (Conv1D)           (None, 8, 128)            49280

 global_max_pooling1d (Globa  (None, 128)              0
 lMaxPooling1D)

 dense (Dense)               (None, 128)               16512

 dense_1 (Dense)             (None, 25)                3225

=================================================================
Total params: 2,514,225
Trainable params: 233,625
Non-trainable params: 2,280,600
_________________________________________________________________
None
```

Figure 8: CNN Model Structure

5 Implementation

The trained kNN, Linear SVM and CNN classifiers are applied to the labelled scaped dataset, which contains gender information. The predicted job titles are evaluated by comparing the gender splits for particular job titles across the three classifiers. The output produced is a predicted job category from each classifier.

By matching the resumes on the unlabelled dataset to the job descriptions scores are produced for each job description.

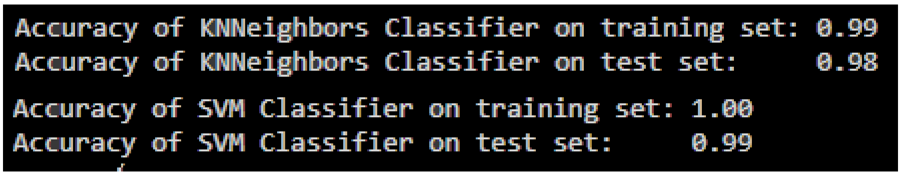
6 Evaluation

The section contains the evaluation of the classifier models and the evaluation of the overall results when the classifiers are applied to the labelled scraped dataset. The results from matching the job descriptions with the resumes from the scraped un-labelled dataset are also evaluated.

[2] https://nlp.stanford.edu/projects/glove/

6.1 Models Evaluation

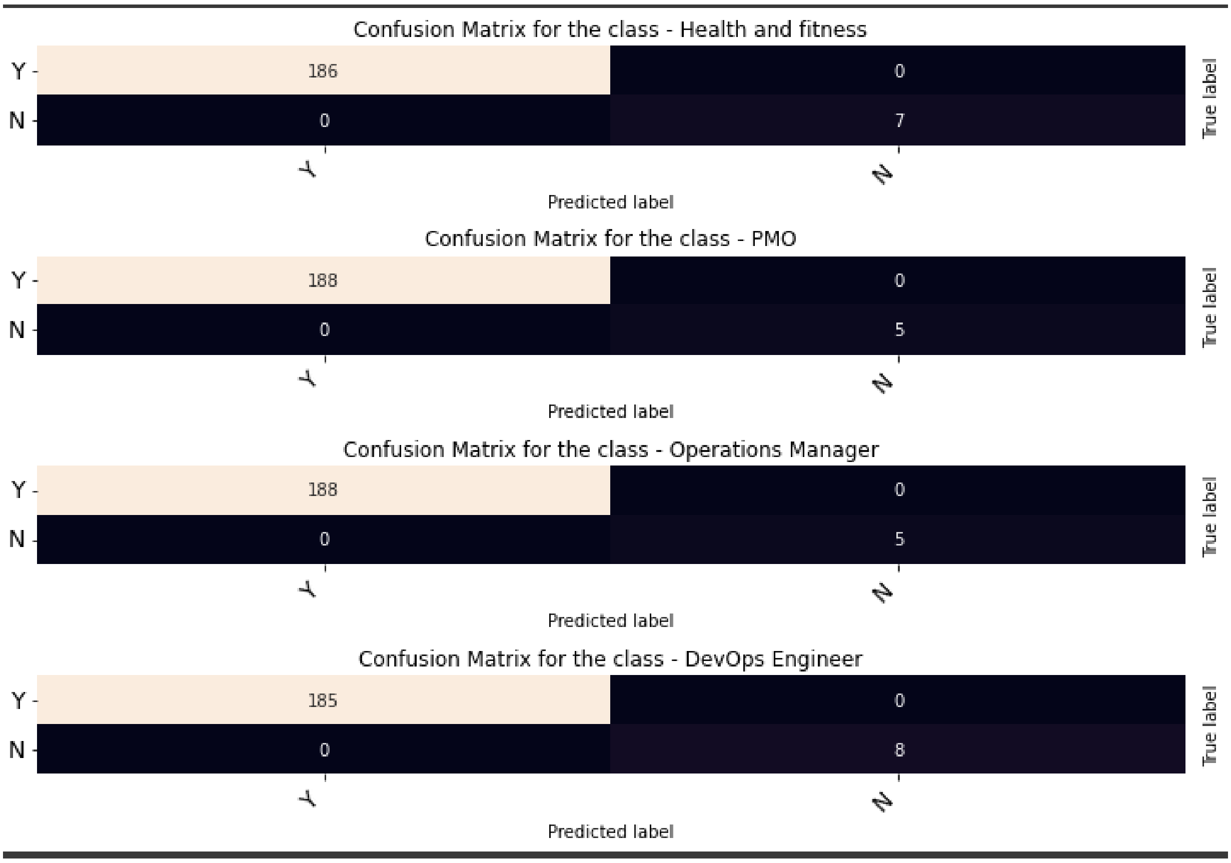
The Linear SVM and kNN classifiers have achieved very high accuracy rates. Figure 9 contains the accuracy rates of the SVM and kNN clasiffiers. However, it is important to consider that due to difficulties in finding a labelled resume dataset, the dataset used is not sufficiently large to provide robust results.



```
Accuracy of KNNeighbors Classifier on training set: 0.99
Accuracy of KNNeighbors Classifier on test set:     0.98

Accuracy of SVM Classifier on training set: 1.00
Accuracy of SVM Classifier on test set:     0.99
```

Figure 9: Accuracy of linear SVM and kNN models



Confusion Matrix for the class - Health and fitness

Confusion Matrix for the class - PMO

Confusion Matrix for the class - Operations Manager

Confusion Matrix for the class - DevOps Engineer

Figure 10: kNN model Confusion Matrix

The evaluation of the CNN model is shown in Figure 11. Accuracy and validation accuracy are used to evaluate the fit of the model. There is a significant difference between these two parameters and this is an indication that the model is over-fitting. This could be due to the size of the dataset as discussed above or due to the combination of layers employed in training the CNN.

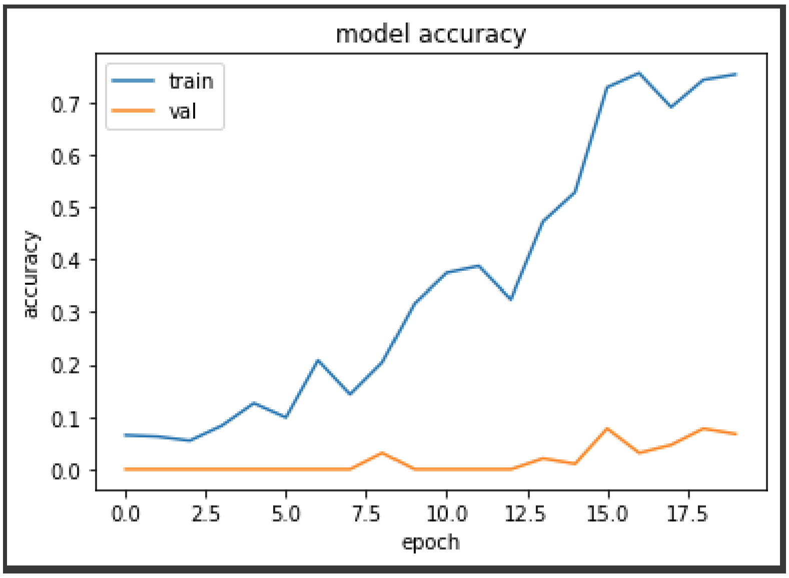


Figure 11: Accuracy of the CNN model vs epochs

6.2 Resumes and Job Description Matching

Figure 12 shows boxplots for a number of job description and the split by gender. The scoring model appears to score the resumes fairly consistently between female and male applicants.

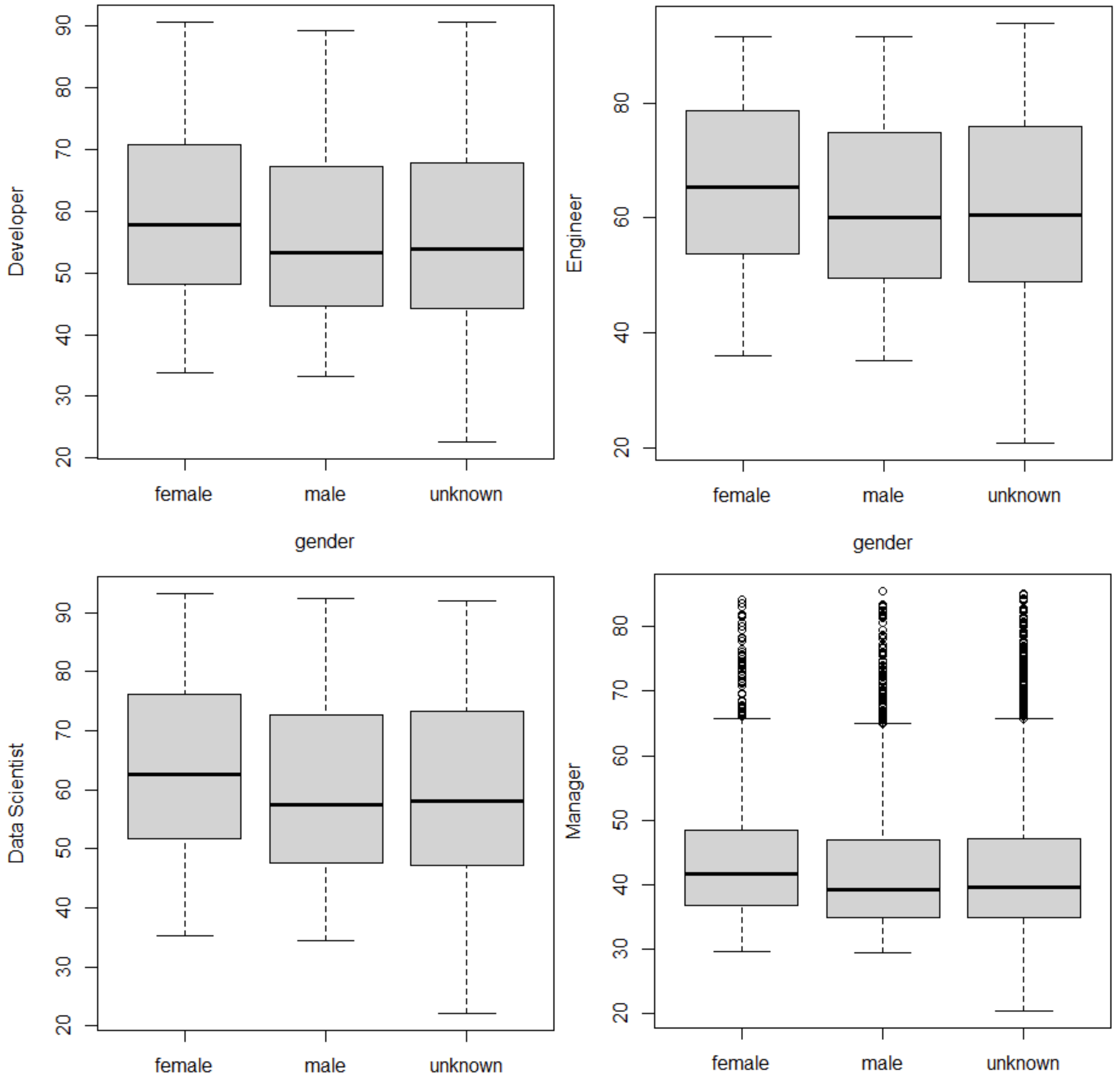


Figure 12: Boxplots for Job Descriptions by Gender

6.3 Classifiers Matching Results

The output of the KNN and Linear SVM classifiers for each job title split by the gender percentage are shown in Figure 14. The scraped labelled dataset has a 60/40 split between men and women. A similar proportion might be expected for the job titles. However, differences are prevalent, for example the Business Analyst job title is predominantly male in the KNN prodiction where it is overwhelmingly female for the SVM classifier, while the CNN classifier shows a fairer split of nearly 50/50. Such discrepancies are concerning and indicate the existence of a bias in the classifiers. Figure 13 contains the gender splits of the Business Analyst role for each of the three classifiers used.

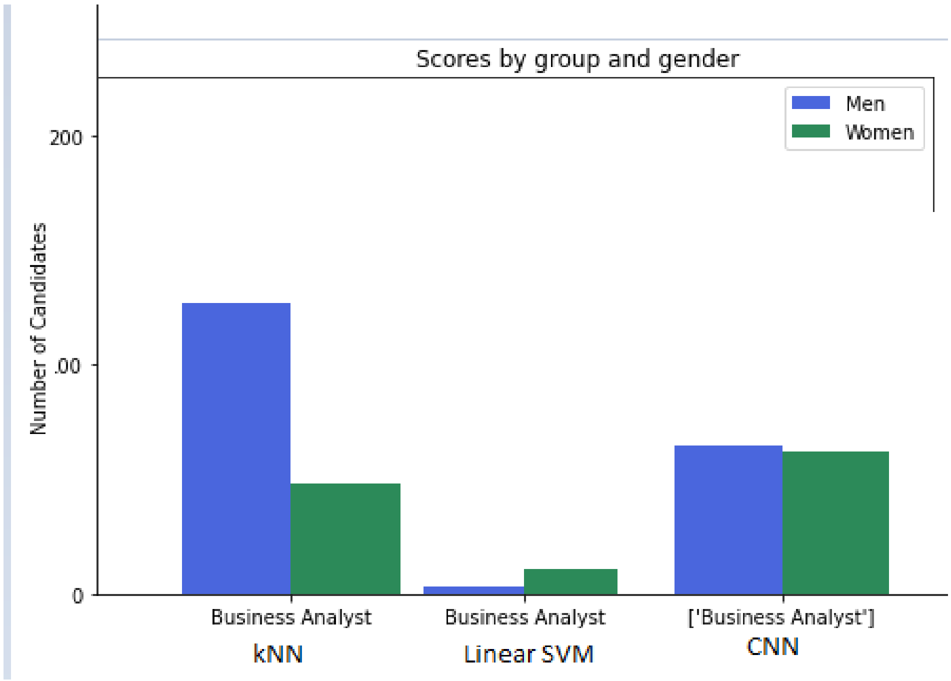


Figure 13: Business Analyst Gender Percentage by Classifier

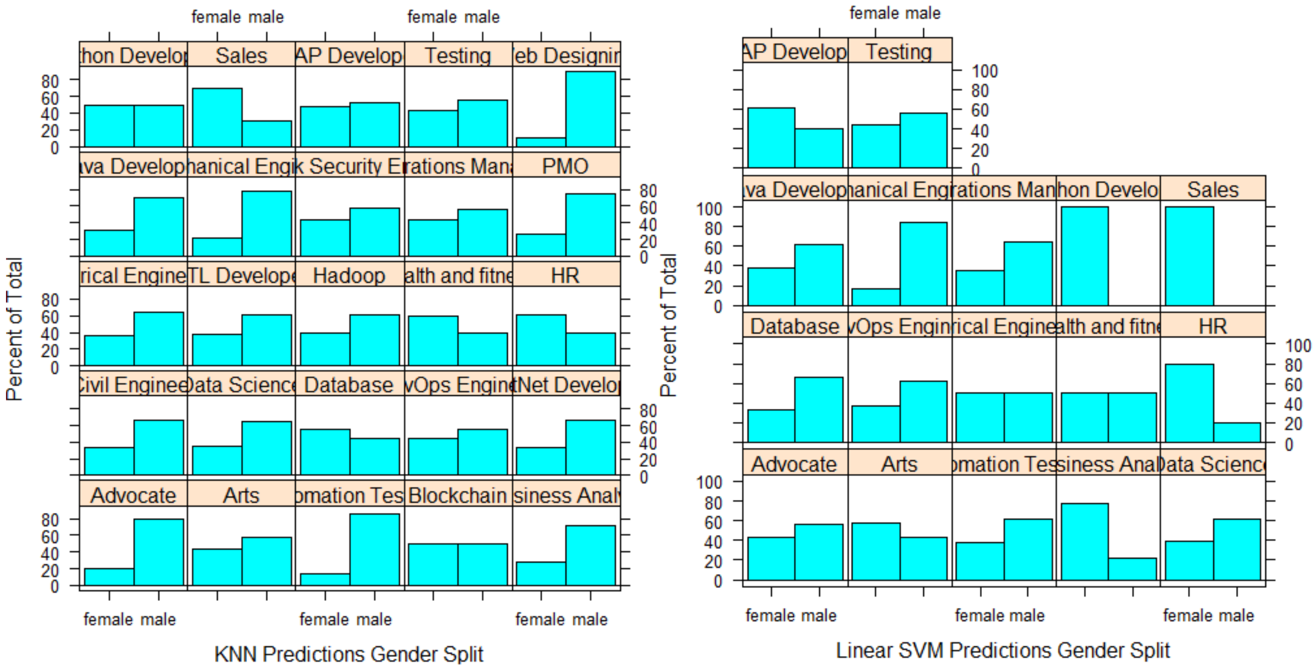


Figure 14: KNN/Linear SVM Classifiers Percentage Gender Split

The results from the CNN predictions are shown in Figure 15

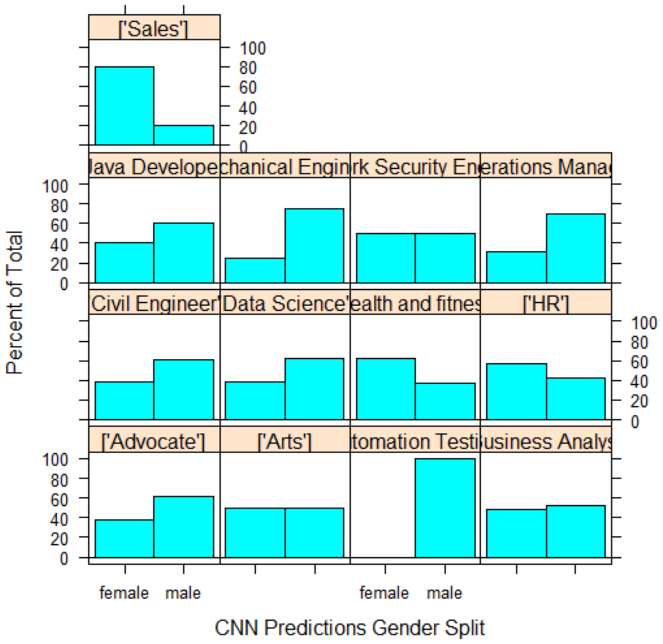


Figure 15: CNN Classifiers Percentage Gender Split

6.4 Discussion

The answer to the research question is that the CNN model does not outperform simpler Linear SVM and kNN models. In fact, its accuracy is significantly lower than SVM and kNN. In addition, the CNN model does not attribute any candidates to a number of categories, which may leave companies struggling to find talent. The best model seems to be the simplest one which measures the text distance as it assigns a score for each job description and may allow recruiters to manually scan CVs with the highest scores.

Due to the limited scope of the research paper the focus has been on evaluating gender bias rather than some more prevalent areas of bias in AI/ML algorithm, such as race, age and ethnic/national origin. However, the research paper provides a blueprint for future research and attempts to demonstrate that some algorithms may be unintentionally bias towards gender due to an incorporated bias in the data on which the dataset is trained.

Additional limitations in the research include the small dataset used to train the models. Scraping, cleaning and appropriately labeling a sufficiently large dataset to train these models has proven to be rather challenging and is beyond the scope of this paper. However, the models used in this paper can be trained in the future at a time when such a dataset becomes available. This will result in more robust predictions.

The experiments conducted indicate that classifiers which are trained on the same data will produce different results when applied to a randomly gathered dataset. The results can be very skewed in terms of gender, for example the Linear SVM model produces exclusively female candidates for the Sales and Python Developer positions. The kNN provides a balanced gender split for those positions. However, the CNN model does an 80/20 split for females and doesn't attribute any resumes to Pyhon Developers. This indicates the importance of having robust models, as otherwise this can lead to discrimination based on gender. The experiments indicate the importance of having robust models which are trained on sufficiently large data-sets. In addition, gender or other discriminatory information should be taken into account building the model so that the model can be tested to ensure that it does not discriminate based on those variables.

A more robust methodology will focus on extracting skill-sets from CVs by training the models to recognize word combinations which indicate a specific skill-set. This is an approach that has widely been used in the literature (Jiechieu and Tsopze; 2021). Collecting a custom datasets allows for more accurate testing of the functionality of the algorithms as the collected datasets simulate real-world CVs.

7 Conclusion and Future Work

Artificial Intelligence and Machine Learning algorithms can be ”black boxes” when it comes to understanding the way decisions are made. Resume recommender systems used by companies are also akin to ”black boxes” in that a company is not accountable to an applicant as to why and how decisions are made. There is no accountability for the way that algorithms operate. As the research has shown different classifiers provide widely different results. This may lead to unconscious bias which is incorporated in the training of the model. Subsequently this bias is passed on in production, where applicants can suffer as a result of the embedded bias. This study demonstrated that sufficiently large dataset is required for training of these models. In addition, the labels with which the classifiers are trained need to be very carefully evaluated to ensure that there are no errors which ultimately can lead to erroneous output from the classifiers. The study showed that the CNN did not perform better than simpler models such as kNN and linear SVM. Simpler techniques, such as the text distance matching applied to the scraped un-labelled dataset proved to be the least biased towards gender out of all the classifiers. However, the study provides significant opportunities for future studies to create a better CNN model which employs alternative layers for training and is trained on a larger dataset to improve results. This research will benefit from a longer timeframe, which will allow the researchers to carefully collect and create the training dataset, understand the data and as a result construct improved models. Future studies can be conducted by including discriminatory parameters such as gender, sex, sexual orientation, national origin, ethnic background etc. as a parameter while building the model. This will ensure that these parameters are being taken into account when making the selection.

References

Alelyani, S. (2021). Detection and evaluation of machine learning bias, *Applied Sciences* **11**(14): 6271.

Ali, I., Mughal, N., Khand, Z. H., Ahmed, J. and Mujtaba, G. (2022). Resume classification system using natural language processing and machine learning techniques, *Mehran University Research Journal Of Engineering & Technology* **41**(1): 65–79.

Bafna, P., Shirwaikar, S. and Pramod, D. (2019). Task recommender system using semantic clustering to identify the right personnel, *VINE Journal of Information and Knowledge Management Systems* .

Calders, T. and Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures, *Discrimination and privacy in the information society*, Springer, pp. 43–57.

Esmaeilzadeh, M., Abdollahi, B., Ganjali, A. and Hasanpoor, A. (2016). Evaluation of employee profiles using a hybrid clustering and optimization model: practical study, *International Journal of Intelligent Computing and Cybernetics* .

Gaur, B., Saluja, G. S., Sivakumar, H. B. and Singh, S. (2021). Semi-supervised deep learning based named entity recognition model to parse education section of resumes, *Neural Computing and Applications* **33**(11): 5705–5718.

Gianfrancesco, M. A., Tamang, S., Yazdany, J. and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data, *JAMA internal medicine* **178**(11): 1544–1547.

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R. and Tomczak, D. L. (2019). "where's the io?" artificial intelligence and machine learning in talent management systems, *Personnel Assessment and Decisions* **5**(3): 5.

Gopalakrishna, S. T. and Vijayaraghavan, V. (2019). Automated tool for resume classification using sementic analysis, *International Journal of Artificial Intelligence and Applications (IJAIA)* **10**(1).

Ilić, S., Marrese-Taylor, E., Balazs, J. A. and Matsuo, Y. (2018). Deep contextualized word representations for detecting sarcasm and irony, *arXiv preprint arXiv:1809.09795* .

Jacovi, A., Shalom, O. S. and Goldberg, Y. (2018). Understanding convolutional neural networks for text classification, *arXiv:1809.08037* .

Jiechieu, K. F. F. and Tsopze, N. (2021). Skills prediction based on multi-label resume classification using cnn with model predictions explanation, *Neural Computing and Applications* **33**(10): 5069–5087.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification, *Artificial Intelligence Review* **52**(1): 273–292.

Landers, R. N. and Behrend, T. S. (2022). Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models., *American Psychologist* .

Leventhal, G. S., Karuza, J. and Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences, *Justice and social interaction* **3**(1): 167–218.

Lin, Y., Lei, H., Addo, P. C. and Li, X. (2016). Machine learned resume-job matching solution, *arXiv preprint arXiv:1607.07657* .

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Mridha, M., Basri, R., Monowar, M. M. and Hamid, M. A. (2021). A machine learning approach for screening individual's job profile using convolutional neural network, *2021 International Conference on Science Contemporary Technologies (ICSCT)*, pp. 1–6.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? a field experiment with thirteen thousand resumes, *American Economic Journal: Economic Policy* **3**(4): 148–71.

Pan, Q., Dong, H., Wang, Y., Cai, Z. and Zhang, L. (2019). Recommendation of crowdsourcing tasks based on word2vec semantic tags, *Wireless Communications and Mobile Computing* **2019**.

Robert, L. P., Pierce, C., Marquis, L., Kim, S. and Alahmad, R. (2020). Designing fair ai for managing employees in organizations: a review, critique, and design agenda, *Human–Computer Interaction* **35**(5-6): 545–575.

Roy, P. K., Chowdhary, S. S. and Bhatia, R. (2020). A machine learning approach for automation of resume recommendation system, *Procedia Computer Science* **167**: 2318–2327.

Trstenjak, B., Mikac, S. and Donko, D. (2014). Knn with tf-idf based framework for text categorization, *Procedia Engineering* **69**: 1356–1364.

Wang, R. F. (2018). Semantic text matching using convolutional neural networks.

Wang, R. and Shi, Y. (2022). Research on application of article recommendation algorithm based on word2vec and tfidf, *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, IEEE, pp. 454–457.

Wings, I., Nanda, R. and Adebayo, K. J. (2021). A context-aware approach for extracting hard and soft skills, *Procedia Computer Science* **193**: 163–172.

Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A. and Kulkarni, D. (2018). Employee attrition prediction, *arXiv preprint arXiv:1806.10480* .

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making, *Data Mining and Knowledge Discovery* **31**(4): 1060–1089.