Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16552-x

# Evaluating and mitigating gender bias in machine learning based resume filtering

**Gagandeep[1] · Jaskirat Kaur[2] · Sanket Mathur[1] · Sukhpreet Kaur[1] · Anand Nayyar[3] · Simar Preet Singh[4] · Sandeep Mathur[5]**

Received: 12 October 2022 / Revised: 1 August 2023 / Accepted: 18 August 2023


**Abstract**
Shortlisting resumes for the companies are being automated using artificial intelligence however, training systems to do that incorporate high social biases in the models. Considering the vitality of mitigating gender bias present in society, the research introduces a method for hiding gender specific terms from data, termed as Gender Masking, before finding the similarity with the job requirements. The paper ideates a method of reduction in indulgence of social biases in machine learning based resume filtering algorithms. In addition, an evaluation method is proposed to justify exclusion of gender specific terms from classification of resumes short-listed for a particular role based upon requirements. The novelty of the proposed method is that upon extraction of information from the resume based on probabilistic indexing, the gender specific terms are masked. This corpus is used as the received information in form of word encoding, across the stated requirements in order to retrieve a similarity score of the information using cosine similarity in correspondence to the posting. The proposed model is evaluated using gender-swapped corpus to ensure unbiased performance of the algorithm. The evaluation method represents the performance variation of the text on swapping the gender, it represents the unintentional differences the algorithm captures based on the biases present in the society. The experimental research is taken out on preprocessed datasets (Online Resume Datasets), from which an average of 15.46% are observed to have been affected by gender bias, which is omitted through the proposed method. From the results computed, an average increase of 1.2% accuracy on the trained Random Forest model is experienced outperforming state-of-the-art techniques of training generic Linear SVM, Logistic Regression and Multinomial Naive Bayes models. The model is regularized to have 100 maximum trees in the ensemble along with 20 maximum depth and 10 minimum samples to split the nodes.

**Keywords** Gender Bias · Information extraction · Resume filtering · Resume classification · Word Embeddings · Vectorization · Gender masking

Extended author information available on the last page of the article

Published online: 01 September 2023

## 1 Introduction

Resume filtering algorithms are used to classify most suitable resumes from all the submitted resumes for the respective opening in the organization. Such filtering systems are implemented using machine learning to find the probabilistic index and similarity index of each applicant to qualify for the requirements [1]. As per LinkedIn, India has the highest proportion of actively job seeking workforce [2] and implementation of such a system has become a necessity for organizations due to the abundance of applicants for the posting and inefficiency of manually performing the task of classification of resumes [3]. A high amount of low-quality and duplicate data is introduced into the system because of the manual interference encountered during collection of data from various sources and through various mediums [5]. Such systems are used to analyze and review the biodatas [4]. However, one drawback of such systems has been incorporation of human biases which restrict the shortlisting of several deserving candidates for job opportunities based on unrelated factors [6].

Machine Learning models are known to be trained on existing historical data in order to predict a value or class of unseen data. The existing data causes the model to capture human bias which is amplified by the parameters tuned by the algorithms [7]. The algorithm captures the existence of gender specific terms together with words representing occupations, performance metrics, achievements etc. in the word embeddings, trained on co-occurrence of words in the text representing groups of words as a n-dimensional word vector [8], to settle a bias towards such terms. Each relationship is characterized by a relation-specific vector offset allowing vector-oriented reasoning based on the offsets between words; these relation-specific vectors when provided to neural networks are observed to be able to contain respective relations in the sentences [9]. Due to this reason, even balanced datasets containing each label co-occur equally display biased predictions towards social aspects as if the dataset had not been balanced [10]. While a machine learning algorithm is trained on a corpus consisting societal biases and stereotypes, the parameters thus learned are observed to have an amplified dependence on these biases, as presented in the dataset, using the principle component analysis to observe variance displayed by eigenvectors and thus it is extremely vital to identify such associations and eliminate them from the system [11, 12, 39]. Such biases have caused companies to take down their filtering systems in the past and have caused unfair evaluation of skills of several deserving candidates over time. Bolukbasi at. el. [12] (2016) discovered that even the word embeddings trained on Google News articles show a high level of gender stereotyping which led to such bias systems. Using linearly separable gender-neutral words, the corpus can be modified for reduction of gender bias into the systems. Aspects like resume writing and socio-linguistics are a source of leakage of bias, which is often translated into biased AI algorithms, this bias is further amplified by the algorithms making the filtering systems robustly biased towards one of the genders based on selection criteria's in the past [13].

The approaches and methodologies discussed above has laid the foundation for a novel methodology to eliminate the factor of biases and impurities in resume selection systems.

### 1.1 Objectives of research paper

The objectives of the paper are as follows:

- The primary objective of the intended research is to create a method for elimination of gender bias while automating the shortlisting systems of resumes respective to the job specifications;
- Extraction of the information from the documents provided as a document into a form compatible for training a machine learning algorithm and then preprocessed the data in order to hide the gender specific terms and trained the machine learning algorithms;
- To evaluate the performance based on the difference in performance of the model on the corpus consisting of datapoints composed of swapped gender specific terms respective to the positioning and relations;
- And, to compare the performance of trained machine learning algorithms on accuracy metrics against state-of-the-art techniques being used for filtering resumes.

### 1.2 Organization of Paper

The paper is organized as follows: Section 2 highlights literature review which iterates over the state-of-the-art methods of performing the processes of creating resume filtering systems as well as the progression towards mitigation of social biases in these systems. Section 3 states the materials used for this research, as well as the methods and techniques used as a base to derive the proposed methodologies. Section 4 enlightens the proposed methodology and functional examples of their working, followed by experimental results, evaluation and comparisons with previous methods in section 5. And, finally, section 6 concludes the paper with future scope.

## 2 Literature review

### 2.1 Case study

Amazon encountered a critical challenge in 2014 with its AI-based resume filtering system when it realized that the algorithms were inadvertently amplifying gender bias during candidate evaluations. Despite aiming to automate the search for top talent, the system's training data, which spanned a 10-year period, primarily comprised resumes from male applicants. As a result, the AI models learned to favor male candidates for technical roles, leading to a biased assessment process. The system displayed unintended discriminatory behavior by penalizing resumes containing gender-specific terms such as "women's" and downgrading graduates from all-women's colleges. This gender bias was a direct reflection of the male-dominated tech industry and the historical data patterns from which the models learned. To address this challenge, Amazon recognized the necessity of ensuring gender neutrality in its hiring process. While the company edited the programs to neutralize specific gender-related terms, concerns remained that the AI could still devise other discriminatory sorting methods.

Ultimately, the project was disbanded by the start of the following year, as executives lost confidence in the tool's ability to overcome gender bias effectively. While recruiters occasionally reviewed the tool's recommendations during candidate searches, Amazon's hiring decisions were not solely reliant on the tool's rankings.

### 2.2 Method review

For the purpose of extraction of information and training models upon it, Lin et al. [1] (2016) and Maheshwari et al. [6] (2010) stated that resume filtering systems are implemented using machine learning to find the probabilistic index of each application to analyze the compatibility with the job requirements, however the drawback of such systems is that they have been considering human bias as a factor for learning the trends. The process of probabilistic indexing consisted of calculating the probability of a given term being gender-specific based on its occurrence and context within the resume dataset. This can be achieved through various methods, such as analyzing the frequency distribution of terms across gender-diverse resumes or using pre-trained language models to estimate the probability of a term being associated with a specific gender. In order to develop a system to filter resumes, the first step is to extract information from the documents in order to build our corpus and word embedding in order to make it compatible to be analyzed by computer systems. Rubenstein and Goodenough [8] and Mikolov et al. [9] (2013) proposed extraction of gender specificity using co-occurrence of words in the embedding. Zhu and Wang [14] and Yu et al. [15] (2005) proposed methodologies stating that the embedding extracted in this step would be representing the information about the application which will be used to find the most suitable applications for the employers' requirements. Chen et al. [17] (2017) proposed such a system of information retrieval from resumes, distinguishing the PDF resumes into two common types viz. Table-styled and List-styled, which as the name suggests are derived from the format of representation of information in the document. Okazaki [16] (2007) used the CRFsuite model, a fast implementation of Conditional Random Fields (CRF) [16], to extract detailed information treating the entire process of information extraction as a sequence labeling problem. In addition, Chen et al. [18] (2018) proposed a Writing Style to distinguish different lines, compared to those extracting algorithms, based on either HMM or CRF, which doesn't need much manually annotated training set. The information retrieved from the document had a level of uncertainty associated with it and the features extracted are thus based on probabilistic terms instead of complete assurance. Van Rijsbergen [19] concluded that the retrieved information is considered to be from an optimal system which manages an efficient trade-off between, first approach, characterizing the document through a representation of its content, regardless of the way in which other documents are described and, second approach, insisting that in characterizing a document one is discriminating it from other documents.

In order to train models for shortlisting, Li et al. [25] (2006) stated that the comparisons between the stills stated in the document to that of requirements need computation of the similarity score. Previous works in this field include, Roy et al. [20] (2020), ideating use of cosine similarity along with k-NN to identify the CVs that are closest to the posted job description working in two steps, i.e., classifying resume into categories and finding out the cosine similarity in respect to the posting; and Zhang et al. [21] (2018) proposed the use of generative adversarial networks by making the generator attempt to prevent discrimination between genders while preventing the discriminator from identifying the gender in a task. Deshpande et al. [13] (2020) stated that socio-linguistics are a source of leakage of bias, this bias is also often translated into AI algorithms. To accompany the findings, Cowgill [36] stated that the model corrects increasingly small distortions and thus small amounts of noise to make good corrections. Guo et al. [7] (2016) and Bolukbasi et al. [12]

(2016) stated that even the bias present in the dataset in small amounts was amplified by the models when passed to them for observing the trends; however, there may exist bias not perceivable directly but only after being observed and amplified by the algorithm getting trained on it. Wang et al. [10] (2019) further expanded the concept of data leakage to be defined as the model inference of irrelevant information as a distinguishing feature, for example, a model inferring that there is likely a woman in an image by identifying a plate in the frame. Wang et al. [29] (2016) examined a deep learning lexical analysis approach proposing a two-channel CNN model to capture features by composing similar and dissimilar components. The system consisted of decomposition of word-vectors using various methods and resulted in an effective way to use ngram word embeddings to find the similarity score among sentences. Chicco [30] proposed another deep learning approach which can be used to find similarity between two input vectors, using the siamese neural networks. These neural networks consist of two identical feedforward neural networks capable of learning representation of an input vector, the results of both these neural networks are then combined and evaluated at the end to find the similarity between the vectors.

The state-of-art methodologies stated techniques to organize the dataset into relation ordinating vectors or swapping the gender-specific terms in the corpus before training the algorithm, resulting in a reduce of gender bias but failing to mitigate it by conserving an imbalance of terms in the dataset. The proposed methodology introduces a preprocessing technique to eliminate gender bias from the data before using it to train the machine learning model or finding the similarity score between the datapoint and the posted job requirement. This is done by masking or hiding gender specific terms with a term specific mask in order to hide the information from the model. This technique deals with modifying the data before analysis and thus is faster than previously stated methods. Unlike traditional methods that simply remove gender-specific terms, this approach considers the surrounding context, allowing the model to retain relevant information while preventing the algorithm from learning gender-related trends. The additional evaluation function ensures mitigation of gender bias from the output generated by the proposed model indicating the need for reiteration of the preprocessing of the dataset. This evaluation method goes beyond traditional evaluation metrics and provides a rigorous analysis of the system's performance in handling gender specificity and bias. The paper emphasizes the ethical significance of introducing controlled noise in the evaluation process to effectively reduce human bias. This consideration is a novel and critical aspect, ensuring the model's fairness and practical deployment in real-world scenarios.

## 3 Material and methods

Section 3 explains existing methods along with examples of how the methods are enhanced and opted for incorporation of our research (Table 1).

### 3.1 Materials

Public dataset of information extracted from a set of resumes is used to evaluate the proposed method in comparison to the previous methods. The dataset properties have been described in Table 2:

**Table 1** Mapping of gender specific terms to masks

| Words | Mask |
|---|---|
| he / she | ** |
| his / her / him | ## |
| himself / herself | ^^ |
| boy / girl | @@ |
| boys / girls | @@* |
| man / woman | && |
| men / women | &&* |

**Table 2** Data details of dataset for comparison of proposed model with previous methods

| Data | Entries | Word Features | Domain |
|---|---|---|---|
| Resume Dataset | 962 | 98,564 | CC0: Public Domain |

URL for download: https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset

**Table 3** Data details for datasets used for evaluation of proposed method

| Data | Entries | Word Features |
|---|---|---|
| DT1 | 311 | 336,855 |
| DT2 | 311 | 313,526 |
| DT3 | 311 | 320,316 |
| DT4 | 311 | 336,692 |
| DT5 | 310 | 310,451 |
| DT6 | 310 | 328,630 |
| DT7 | 310 | 333,730 |
| DT8 | 310 | 331,082 |

URL for download: https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/metadata

The dataset used for further investigation and measuring generalized performance is downloaded from online resources of Kaggle. For convenience of examining the performance of our methodologies on the dataset, we have labeled our dataset at DT1, DT2, DT3, DT4, DT5, DT6, DT7, and DT8. The details about the dataset are described in Table 3. The visualization of bias calculated from the datasets are as represented in Figs. 1, 2 and 3.

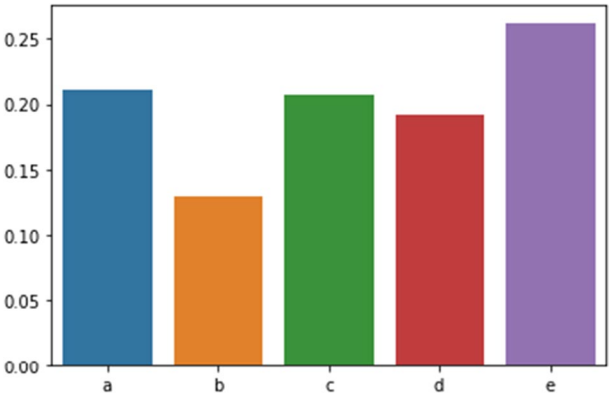


**Fig. 1** Result of similarity score of applicants

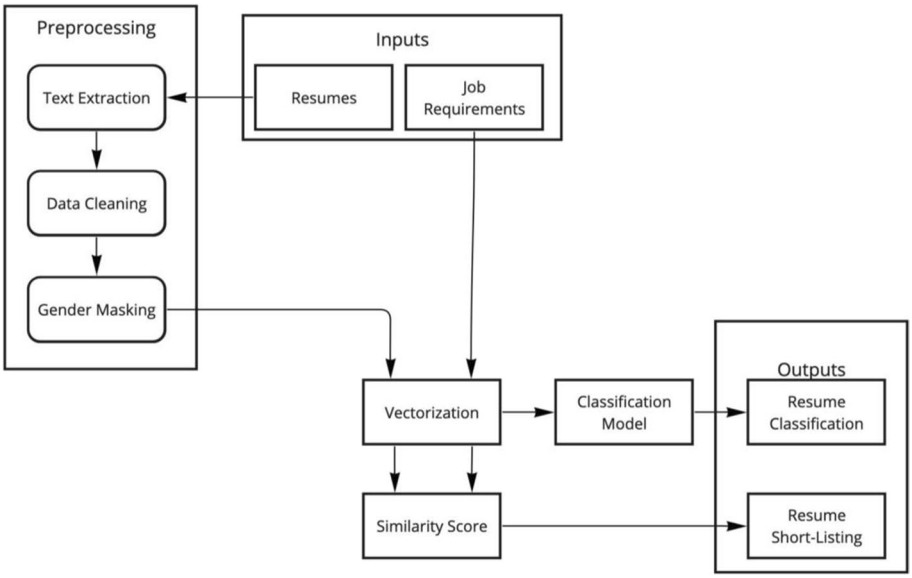


**Fig. 2** Flow chart of entire proposed model

### 3.2 Methods

A common technique used for training models is on a union of data replacing all male entities with female entities and vice versa to ensure that the model identifies that the results are the same for the retrieved information regardless of the gender [22]. The sentences like, "She is a software engineer" is swapped to create another data point, "He is a software engineer" and this process of swapping terms which represent gender of the gendered nouns is referred to as Gender-swapping. It can be used to generalize the sentences and if the model does not make decisions based on genders, it should perform equally for both sentences, otherwise, the difference in evaluation scores reflects the extent of gender bias found in the system [11, 23, 24]. Finding the similarity between the text [25] and requirements can be done in several ways out of which some perform better for different use cases, including some prewritten techniques in the packages. Python packages SimMetrics,

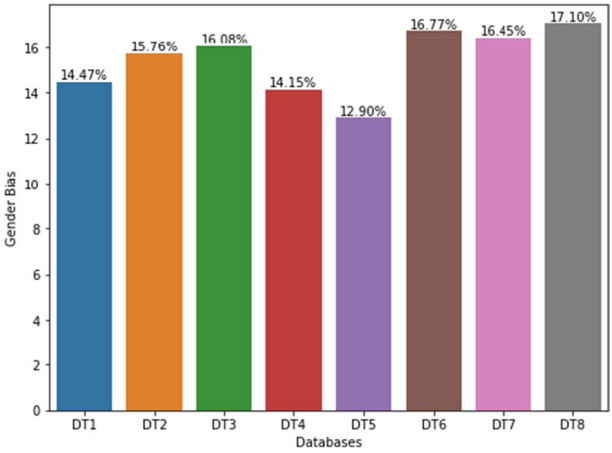


**Fig. 3** Gender Bias present in the Datasets

WordNet. Similarity and NLTK were found to be a few prewritten algorithms, along with several techniques like string algorithms and training models to find similarity between text effectively [26–28, 40, 41].

The semantic similarity has been derived using the following base researched methods introduced in 2016 [37]. The decomposition function adapts to semantic matching expression by determining whether there is an exactly matched word in the other sentence, as represented in Eq. (1) (Adopted from Wang et al. [29]).

$$\begin{bmatrix} s_i^+ = s_i; s_i^- = 0 \\ s_i^+ = 0; s_i^- = s_i \end{bmatrix} \begin{matrix} if\ s_i = \hat{s}_i \\ otherwise \end{matrix} \tag{1}$$

where,

$s_i$ word vector component
$\hat{s}_i$ semantic vector component
$s_i^+$ similar components
$s_i^-$ dissimilar components

Which is then used to find the similarity between the vectors and is represented by the cosine similarity of the vectors. This has been represented in the Eq. (2) (Adopted from Wang et al. [29]) along with the use of coefficients with respect to finding the similarity.

$$\begin{aligned} \alpha &= \frac{s_i^T \hat{s}_i}{\|s_i\| \cdot \|\hat{s}_i\|} \\ s_i^+ &= \alpha s_i \\ s_i^- &= (1-\alpha) s_i \end{aligned} \tag{2}$$

where,

$\alpha$ cosine similarity measure
$s_i^T$ transpose of the word vector components

Word embedding prepared for the scope of this paper, is using TF-IDF or Term Frequency - Inverse Document Frequency as represented in Eq. (3) (Adopted from Christian et al. [42]), from the corpus containing trigrams, which are prepared from 3 consecutive pairs of words from the text corpus used to capture the context of words in each document.

$$TF - IDF = \left(\frac{w_d}{w_t}\right) \times \log\left(\frac{N}{d_w}\right) \tag{3}$$

where,

| | |
|---|---|
| $w_d$ | count of word occurrence in the document |
| $w_t$ | total words in document |
| $d_w$ | number of document the word is in |
| N | total number of documents |

The similarity of elements for this application is being calculated by cosine similarity, and is measured as the cosine of the angle between two vectors on a plane as given in Eq. (4) (Adopted from Han et al. [38]). It is only measured in positive spaces and thus the value of cosine similarity ranges between 0 to 1, where 0 represents cos(90°) or completely dissimilar vectors, and 1 represents cos(0°) or overlapping vectors.

$$\cos(x, y) = \frac{x\, y\cdot}{\|x\| * \|y\|} \tag{4}$$

where,

| | |
|---|---|
| x, y | represents two vectors |
| x.y | dot product of the two vectors |
| ‖x‖ | magnitude of the vector |

## 4 Proposed methodology

This section includes explanation of the proposed model with respect to use of insights from existing solutions and previous approaches.

### 4.1 System model

The entire framework for the proposed method of training the model on the dataset is preprocessed with Gender-Masking in order to mitigate the gender bias present in the resumes, is represented in Fig. 2.

The input for this process is the job requirements and the information extracted from the resumes submitted by the applicants. The extracted information corpus is then preprocessed in order to clean the dataset along with the proposed method of Gender-Masking. This processed information is then vectorized using TF-IDF and used for shortlisting of applicants. The classification model is used in order to classify the domain according to the resume in order to filter the applicants for the roles offered by

the job posting. The top applicants out of this pool of classified resumes are shortlisted based on the similarity score of the resume with the posted requirements.

The data to be used in this system has been referred to as gender-masked corpus and describes a technique of masking the gender-specific terms which exist in the retrieved information. In order to mitigate the bias due to word embedding of corpus in the form of n-dimensional series of co-existing words, all the gender specific terms are replaced by a mask. Using this technique, any opportunity of data leakage into the model is eliminated, which might have caused it to show a bias. The TF-IDF vectorizer is fitted upon the cleaned and masked data in order to capture the relevant masks from the trigrams. Hiding selective information has been used for various purposes, like protecting content and authentication, concealing the information inside another document is termed as steganography as it is highly used for the purpose of data hiding [31, 32]. Masking information can also be used to make text topic-neutral by masking topic related terms and is used to enhance authorship [33]. A mask in this reference is described as a character or series of characters used to replace any gender specific word in the retrieved text. Replacement of the character ensures retrieval of correct information embedding without causing the series to have missing words and combining unrelated words together for identification by the algorithm.

### 4.2 Functioning of gender-masking

Gender-Masking is the proposed method, representing hiding gender specific information from the corpus. Following are the examples of the gender-masking process:

Example 1:

Original Text Retrieved: *Improved search time of queries in the Men's Apparel section by 8.2% with queries up to 50 characters.*

Gender-Masked Text: *improved search time of queries in the &&*'s apparel section by 8.2% with queries up to 50 characters.*

Example 2:

Original Text Retrieved: *Consulted a women led tech startup to re-design their business model to increase the profit by 34% and market share by 12.8% while she was expanding her business in India.*

Gender-Masked Text: *consulted a &&* led tech startup to re-design their business model to increase the profit by 34% and market share by 12.8% while ** was expanding ## business in India.*

The masking is performed based on a predetermined mapping of gender-specific terms and their mapping to a mask. The mapping of the gender-specific terms to the mask is defined as mapping of word to a representation of a sequence for all pronouns and words with the same indulgence. The sample mapping used in the above example is given in Table 1.

### 4.3 Similarity between applications and posting

In order to assign a similarity score to the applications based on the information on their resumes, we have used cosine similarity on the word embedding extracting after the gender-masking and preprocessing procedures. The score will be a relative metric for our use-case in order to shortlist top 'n' candidates from the pool of all applications. The score can be visualized as the distance between the embedding and the stated requirements. The

similarity score will be a conversion of the cosine similarity score of each datapoint to be included in a group to a relative metric across the stated requirements [34, 35]. An application whose information has a datapoint near all of the requirements will represent a skill set matching the posting necessities and expected skills and thus will be assigned a higher score than those applicants whose skills lie further apart.

For the output of the proposed algorithm, we need to find the independent score of each applicant in order to find the relevant scores from each one of them. In order to achieve this, normalization of each score in respect to other to short-list top 'n' candidates has been performed.

$$Score(x_i) = \frac{\cos(x_i, \quad req)}{\sum_{j=0}^{n} \cos(x_j, \quad req)} \tag{5}$$

where,

| | |
|---|---|
| $x_i$ | $i^{th}$ vector from the list of vectors x |
| req | vector of requirements posted for job application |
| cos(a,b) | cosine similarity defined in Eq. (4) |

### 4.4 Technical functionality of similarity score

Following is the example of finding relative similarity score of applicants with the requirements in the job posting:

Requirement:*A data scientist with skills in Python, NumPy stack, TensorFlow, Google Cloud and optionally R.*

Applicants:

a: *3 years of experience with Python (NumPy, pandas, TensorFlow, sklearn,* etc.*)*

b: *Working as a web developer and DevOps engineer with google cloud and azure.*

c: *Handling data science projects with R along with deployment to Google Cloud and AWS.*

d: *Holding a position of Junior Data Scientist at a company that is growing rapidly.*

e: *A data scientist with strong experience in python, R and TensorFlow.*

**Results** As represented in Fig. 1, applicant 'e' was considered as the most suitable candidate for the role according to the job requirement.

## 5 Experimentation, results and analysis

### 5.1 Experimental setup

The research is taken out on a system with intel core i5 9th generation processing unit along with NVidia GeForce GTX 1650 graphic card of 4GB capacity. The disk architecture contains a 256 GB solid state drive along with 1 TB Hard Drive. In addition, Kaggle notebook and google Colab cloud architectures were used for processing data and training algorithms through targeted compute components.

### 5.2 Experimental parameters and measurements

The figures in the following sections represent the performance of various methods on the databases. The methods are named as,

- Method 1 - Linear Support Vector Machine Classifier
- Method 2 - Random Forest
- Method 3 - Multinomial Naive Bayes
- Method 4 - Logistic Regression
- PM - Proposed Method (Random Forest on Gender-Masked dataset)

In order to incorporate the assurance of mitigation of gender bias in the data, we are proposing a custom evaluation function using gender-swapped information as the data points. The cosine similarity of each document is measured with the gender-swapped data points in each iteration and will be used to evaluate the existence of gender-specificity in the corpus which might lead to a gender bias when used to train a machine learning algorithm. As discovered by Deshpande et al., [13], in 2020, the bias screeners must be biased and sufficiently noisy in order to reduce human bias, reduction in noise will worsen the model performance and its ability to correct human bias. As the amount of noise in human decisions increases, the model corrects increasingly small distortions and thus it only needs a small amount of noise [36].

Using the evaluation function, it is observed that swapping the gender-specific terms caused 7.831% of documents from the dataset being used, consisting of gender-specificity associated with them using the defined dictionary. The gender-masking can be performed at a larger scale with various additions to the gender-specific terms in the dictionary we generate depending upon the use case. The methodology of the evaluation function used for finding gender bias (GB) in a document is described in Eq. (4).

$$GB = \frac{count\left(\left\{Score(v, \quad req) \neq Score\left(v', \quad req\right), \quad v \in x\right\}\right)}{N} \times 100 \tag{5}$$

where,

| | |
|---|---|
| GB | gender bias in the filtering method. |
| x | list of vectors. |
| v’ | vectors with gender-swapped terms. |
| N | total number of vectors in the list. |
| Score(a,b) | scoring function defined in Eq. (3). |

The GB evaluation metric is then used to equate the gender specificity existing in the dataset which may cause gender leakage into the system and result in a biased model for filtering job applicants based on their gender.

### 5.3 Results and comparisons on the dataset

We calculated the impact of gender specific on the datasets of the resumes, and these observations reflect that the masking of gender specific terms does have an adverse effect

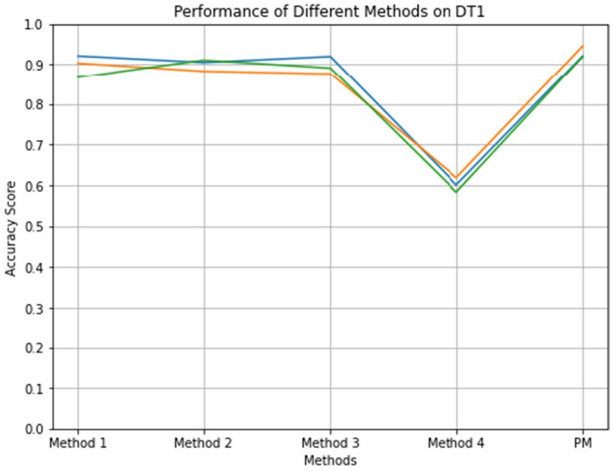


**Fig. 4** Performance of various methods on DT1 with different random splitting

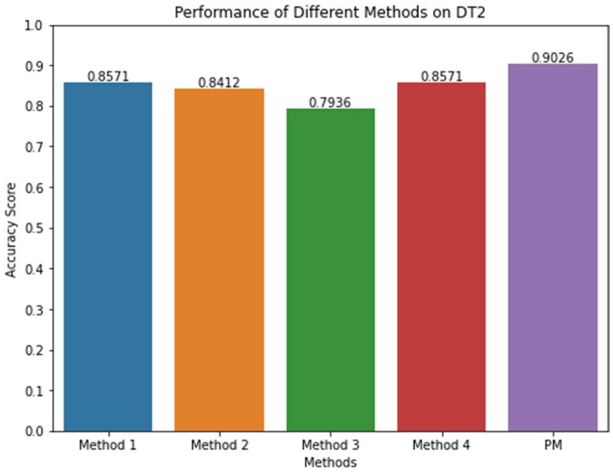


**Fig. 5** Performance of various methods on DT2

on the machine learning models trained and by rationality, deduce the gender bias from these algorithms. Testing the proposed model on various datasets, we have found the existence of gender bias using our evaluation function stated in Eq. (4).

The consistently exposure of the existence of gender bias is reinforcing the necessity of our approach to mitigate biases in automated resume filtering algorithms.

Figure 4 examines execution of the implementation of our method on dataset DT1. The various lines in the graph represent performance of the model on different random splitting of the dataset for training and evaluating accuracy of the methods. From the three random splitting, the proposed method evaluated to be equivalent or improved in all the cases.

The proposed method when evaluated on dataset DT2 displayed a 0.0455 improvement in the accuracy metrics as compared to the previous methods which were being executed without applying Gender Masking while preprocessing the corpus, as represented in Fig. 5. Likewise, the evaluation on dataset DT3, as depicted in Fig. 6, demonstrated a subtle yet

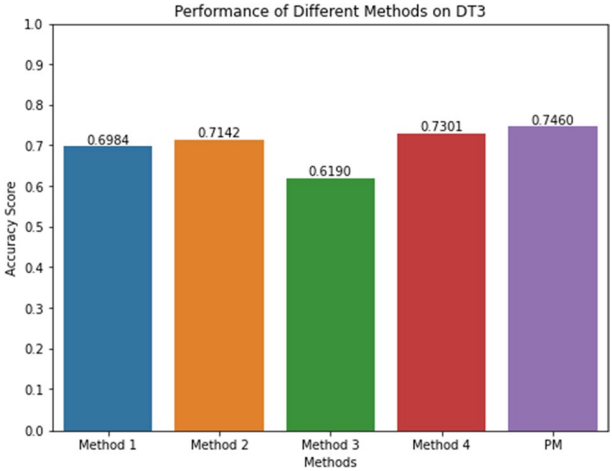


**Fig. 6** Performance of various methods on DT3

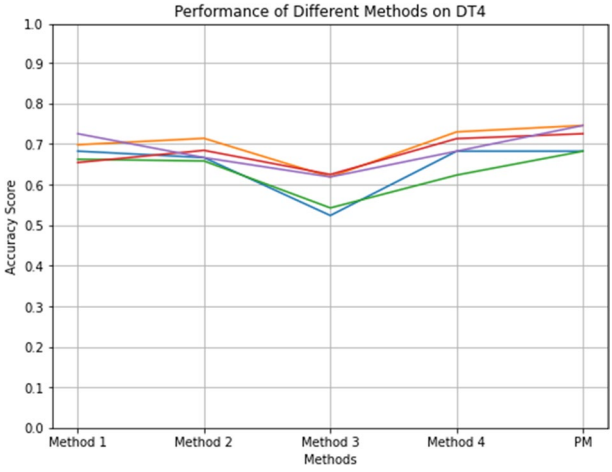


**Fig. 7** Performance of various methods on DT4 with different random splitting

consistent increase in the evaluation metric for the proposed method, further validating its effectiveness in mitigating gender bias in the automated resume shortlisting process.

Figure 7 visualizes performance of the methods on 5 random splitting of the dataset represented by each line in the graph. The method proposed in this paper evaluated to be better performing. The method also eliminates the gender bias from the dataset, making the model perform better as well as be unbiased while filtering applicants based on their gender.

The dataset DT5 is evaluated in 5 partitions of the entire dataset. The data is distributed into the partitions uniformly and used individually in order to train the model on the preprocessed data. The performance of the proposed method is observed to be better than other methods in all the partitions of the dataset. The results of the evaluations of the dataset are visualized in Fig. 8.

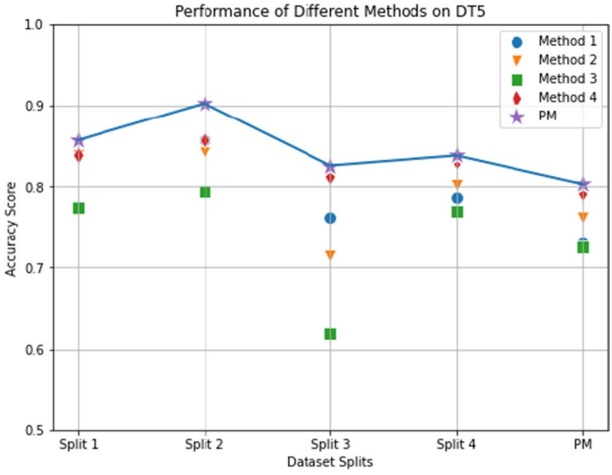


**Fig. 8** Performance of various methods on DT5 with divisions of dataset into separate splits

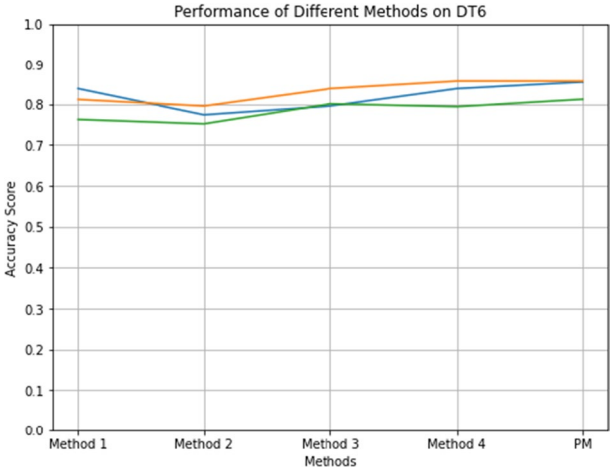


**Fig. 9** Performance of various methods on DT6

Figure 9 represents evaluation of dataset DT6 on various methods for comparison done through different splitting of data into the training and testing sets. The proposed model performed equivalent or better in all different splitting of the dataset.

The performance of proposed model on dataset DT7 is slightly lower than achieved with previous model as represented in Fig. 10 and same is the condition with performance on split 4 partitioned from DT8 as visualized in Fig. 11, however, rest 4 splits performed better with proposed model.

Finally, Fig. 12 visualizes the performance of the proposed method compared to that of the previous methods on various datasets.

The comparison clearly shows that the performance of machine learning models have increased when using Gender-Masking as a preprocessing step. The method also ensures mitigation of gender bias from the filtering systems based on the resumes for a provided job description.

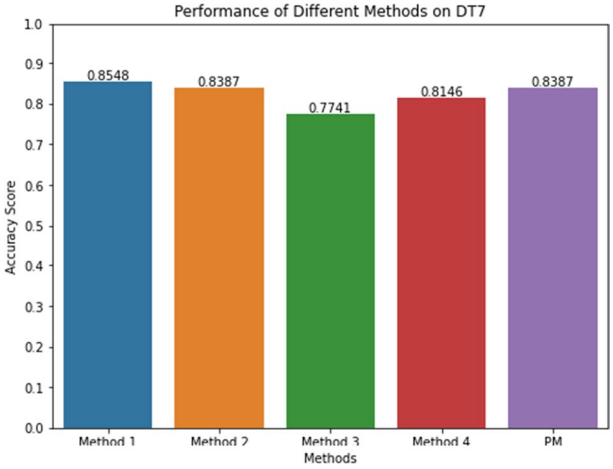


**Fig. 10** Performance of various methods on DT7

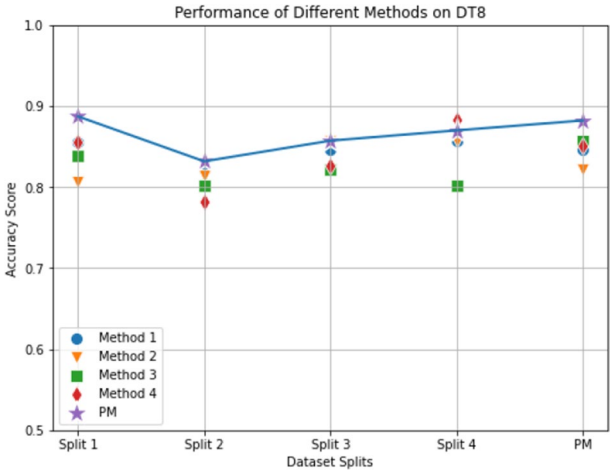


**Fig. 11** Performance of various methods on DT8 with divisions of dataset into separate splits

## 5.4 Experimental analysis

Since the corpus has been preprocessed with cleaning techniques and gender-masking as well as each document assigned to a similarity score according to the job requirements, we will analyze how well a machine learning model can classify the gender-masked documents as resumes of different roles. Roy et al. [20], classified the data extracted from resumes to various job profiles using 4 machine learning models viz. Random Forest, Multinomial Naive Bayes, Logistic Regression and Linear Support Vector Machine Classifier with the results stated in Table 4. According to the findings, the Linear Support Vector Machine Classifier worked the best with an accuracy of 0.7873. Training the above stated model with changed parameters and gender-masked data as proposed in this paper, received results are stated in Table 4. The random forest classifier worked the best after the preprocessing steps with an accuracy of 0.7941. The hyperparameters used while training the model are as follows:

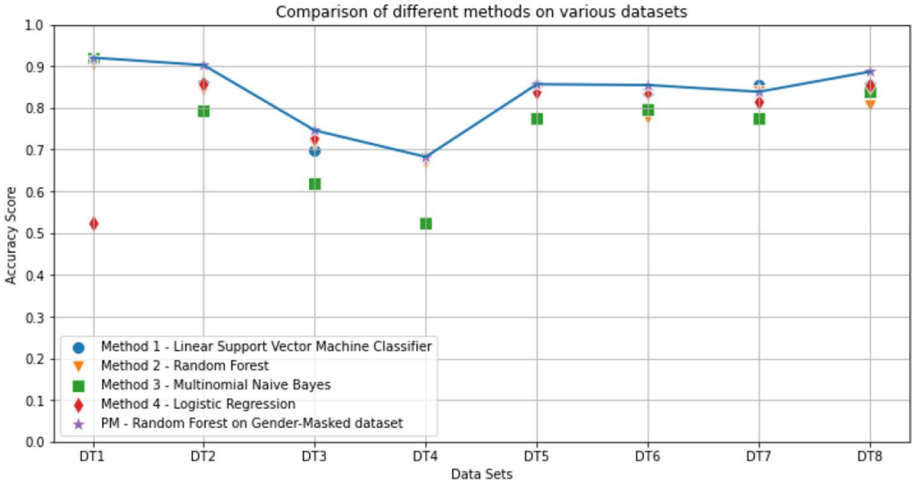


**Fig. 12** Comparison of performance of proposed and previous methods on various datasets (DT1-DT8)

**Table 4** Comparison of accuracy with previously used methods

| Classifiers | Previous Methods | After Proposed Gender Masking Method |
|---|---|---|
| Linear Support Vector Machine Classifier | 0.7853 | 0.5294 |
| Logistic Regression | 0.6240 | 0.6470 |
| Multinomial Naive Bayes | 0.4439 | 0.7647 |
| Random Forest | 0.3899 | **0.7941** |

Bold value is showing the best performance value by the proposed method in the table

- Linear Support Vector Machine Classifier: The regularization parameter is set to the coefficient 100. The strength of the regularization is inversely proportional to the coefficient. This parameter is used to penalize the model based on a squared L2 penalty.
- Logistic Regression: Same as the Linear Support Vector Machine Classifier, 100 is used as the regularization parameter with L2 penalty as default.
- Multinomial Naive Bayes: The additive smoothing parameter is set to 0.01 to ensure smoothing of the model fitted by the algorithm.
- Random Forest: The ensemble method used 100 trees in the forest with 20 as the maximum depth each tree allowed to have. The minimum number of samples required to split a node set to 10 which was decided according to the size of the dataset.

After conducting the comparative study, it is evident that the proposed preprocessing steps, including gender-masking, have a substantial impact on the performance of the machine learning models used for resume classification. Notably, the Random Forest classifier outperformed other models, including the previously top-performing Linear Support

Vector Machine Classifier by a slight margin, achieving an accuracy of 0.7941 compared to the previous 0.7873.

The key takeaway from the comparative study is that the incorporation of gender-masking and other preprocessing techniques proposed in this paper significantly enhance the model's ability to classify gender-masked documents as resumes for different job roles. The Random Forest classifier, with its ensemble approach and carefully tuned hyperparameters, demonstrates a superior understanding of the gender-masked data, resulting in improved accuracy in comparison to the previous models. These findings underscore the importance of addressing gender bias in automated resume classification, as the models trained on gender-masked data deliver more equitable and unbiased results. The increased accuracy of the Random Forest classifier post-preprocessing highlights the potential of the proposed method in promoting fair and inclusive hiring practices by reducing gender bias in the machine learning algorithms' decision-making processes.

## 6 Conclusion and future scope

As big of a concern gender inequality is in society, bigger is the concern about amplification of that bias in the machine learning algorithms which are expected to analyze and shortlist applicants for a job role fairly. Several analyses over time have referenced an unequal representation of males and females in the filtering systems with a bias of preference to male candidates over female candidates for job roles. Using the proposed system, we are progressing towards minimizing any opportunity of leaking gender-specific data to the model training phase in order to restrain it from observing trends related to gender in the corpus and word-embeddings. Mapping the gender-specific terms to a pre-distinguished set of symbols representing the context, as termed as Gender Masking, is helping in generalizing the model towards genders while also confining relations and representations in the sentences. These sentences when converted to word-embeddings represent the information irrespective of genders. Furthermore, using gender-masking in the evaluation function, we are identifying any data leakage or gender specificity in the system, the insights which can be then used to neutralize the bias in the corpus and mitigate the impact on the algorithms that will be training using the dataset. As a result of applying the proposed method to the selected corpus, the preprocessing resulted in an increase of 1.2% accuracy on the trained random forest model over previous techniques.

The prospects for future research would be building a thorough masking dictionary with more gender-specific terms and elaborated masks. The gender-specific terms expand more than what we have used for the purpose of this research, thus providing us an opportunity to include more such terms and expand their mapping to relative masks. The masks can be improved by use of hashing techniques in order to be able to sustain more aspects about the interpretation along with maintaining gender neutrality. Additions to this research would help bring gender ratio equilibrium in the industry, helping with reduction of bias from the system altogether restraining social biases to manipulate the employment sector. While the primary focus is on gender bias, the paper acknowledges the potential for extending the approach to other forms of bias. This insight opens avenues for future research in mitigating biases related to race, ethnicity, and other sensitive attributes.

**Funding** The authors received no specific funding for this study.

**Data availability** Authors declare that all the data being used in the design and production cum layout of the manuscript is declared in the manuscript.

## Declarations

**Conflict of interest** The authors declare that they have no conflicts of interest to report regarding the present study.

## References


1. Lin Y, Lei H, Addo PC, Li X (2016) Machine learned resume-job matching solution. Computation and Language, ArXiv.https://doi.org/10.48550/arXiv.1607.07657
2. Howard JL, Ferris GR (1996) The employment interview context: social and situational influences on interviewer decisions 1. J Appl Soc Psychol 26(2):112–136
3. Zhang L, Fei W, Wang L (2015) Pj matching model of knowledge workers. Procedia Comput Sci 60:1128–1137
4. Breaugh JA (2009) The use of biodata for employee selection: past research and future directions. Hum Resour Manag Rev 19(3):219–231
5. Roy PK, Singh JP, Baabdullah AM, Kizgin H, Rana NP (2018) Identifying reputation collectors in community question answering (CQA) sites: exploring the dark side of social media. Int J Inf Manag 42:25–35
6. Maheshwari S, Sainani A, Reddy PK (2010) An approach to extract special skills to improve the performance of resume selection. In International workshop on databases in networked information systems (pp. 256–273). Springer, Berlin, Heidelberg
7. Guo S, Alamudun F, Hammond T (2016) RésuMatcher: A personalized résumé-job matching system. Expert Syst Appl 60:169–182
8. Rubenstein H, Goodenough JB (1965) Contextual correlates of synonymy. Commun ACM 8(10):627–633
9. Mikolov T, Yih WT, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies, Association for Computational Linguistics, pp 746–751
10. Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V (2019) Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations. In: proceedings of the IEEE/CVF international conference on computer vision, Computer Vision and Pattern Recognition, pp 5310–5319
11. Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, Mirza D, Belding E, Chang K, Wang WY (2019) Mitigating gender bias in natural language processing: Literature review. arXiv. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp 1630–1640
12. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, pp 1–9. https://arxiv.org/abs/1607.06520
13. Deshpande KV, Pan S, Foulds JR (2020) Mitigating demographic Bias in AI-based resume filtering. In: Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization. Adaptation and Personalization. Association for Computing Machinery, pp 268–275. https://doi.org/10.1145/3386392.3399569
14. Zu S, Wang X (2019) Resume information extraction with a novel text block segmentation algorithm. Int J Nat Lang Comput 8:29–48
15. Yu K, Guan G, Zhou M (2005) Resume information extraction with cascaded hybrid model. In: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05), pp 499–506. https://doi.org/10.3115/1219840.1219902
16. Okazaki N (2007) Crfsuite: a fast implementation of conditional random fields (crfs) http://www.chokkan.org/software/crfsuite/
17. Chen J, Gao L, Tang Z (2016) Information extraction from resume documents in pdf format. Electron Imaging 2016(17):1–8
18. Chen J, Zhang C, Niu Z (2018) A two-step resume information extraction algorithm. Math Probl Eng 2018:8. https://doi.org/10.1155/2018/5761287

19. Van Rijsbergen C (1979) Information retrieval: theory and practice. In: Proceedings of the Joint IBM/ University of Newcastle upon Tyne Seminar on Data Base Systems, vol 79, pp 1–14
20. Roy PK, Chowdhary SS, Bhatia R (2020) A machine learning approach for automation of resume recommendation system. Procedia Comput Sci 167:2318–2327
21. Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, Machine learning, pp 335–340. https://doi.org/10.48550/arXiv.1801.07593
22. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2018) Gender bias in coreference resolution: evaluation and debiasing methods. arXiv. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 2, pp 15–20
23. Lu K, Mardziel P, Wu F, Amancharla P, Datta A (2020) Gender bias in neural natural language processing. In Logic, language, and security (pp. 189–202). Springer, Cham
24. Kiritchenko S, Mohammad SM (2018) Examining gender and race bias in two hundred sentiment analysis systems. arXiv. In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (SEM).https://doi.org/10.48550/arXiv.1805.04508
25. Li Y, McLean D, Bandar ZA, O'shea JD, Crockett K (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 18(8):1138–1150
26. Singh S, Singh H, Gehlot A, kaur J, deep G (2023) IR and visible image fusion using DWT and bilateral filter. Microsystem Technologies 29(4):457–467
27. Islam A, Inkpen D (2008) Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans Knowl Discov Data (TKDD) 2(2):1–25
28. Pradhan N, Gyanchandani M, Wadhvani R (2015) A review on text similarity technique used in IR and its application. Int J Comput Appl 120(9):29–34
29. Wang Z, Mi H, Ittycheriah A (2016) Sentence similarity learning by lexical decomposition and composition. arXiv. In: Proceedings of Coling 2016.https://doi.org/10.48550/arXiv.1602.07019
30. Chicco D (2021) Siamese neural networks: An overview. Artificial Neural Networks, vol 2190, pp 73–94. https://doi.org/10.1007/978-1-0716-0826-5_3
31. Bennett K (2004) Linguistic steganography: survey, analysis, and robustness concerns for hiding information in text, Computer Science, Purdue University, 2004
32. Narayana VL, Kumar NA (2018) Different techniques for hiding the text information using text steganography techniques: a survey. Ingénierie des Systèmes d'Information 23(6):115–125
33. Stamatatos E (2018) Masking topic-related information to enhance authorship attribution. J Assoc Inf Sci Technol 69(3):461–473
34. Xia P, Zhang L, Li F (2015) Learning similarity with cosine similarity ensemble. Inf Sci 307:39–52
35. Park K, Hong JS, Kim W (2020) A methodology combining cosine similarity with classifier for text classification. Appl Artif Intell 34(5):396–411
36. Cowgill B (2018) Bias and productivity in humans and algorithms: theory and evidence from resume screening. Columbia Business School, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.2139/ssrn.343373729
37. Celik D (2016) Towards a semantic-based information extraction system for matching résumés to job openings. Turk J Electr Eng Comput Sci 24(1):141–159
38. Han J, Kamber M, Pei J (2012) Getting to know your data. In Data mining (Vol. 2, pp. 39–82). Morgan Kaufmann, Boston, MA
39. Deep G, Kaur J, Singh SP, Nayak SR, Kumar M, Kautish S (2022) MeQryEP: A Texture Based Descriptor for Biomedical Image Retrieval. J Healthc Eng 2022:20. https://doi.org/10.1155/2022/9505229
40. Solanki A, Kumar A, Rohan C, Singh S P, Tayal A (2019) Prediction of breast and lung Cancer, comparative review and analysis using machine learning techniques. In: Smart computing and self-adaptive systems. CRC Press, Boca Raton, pp 251–271

41. Kaur H, Singh S P, Bhatnagar S, Solanki A (2021) Intelligent smart home energy efficiency model using artificial intelligence and internet of things. In: Artificial intelligence to solve pervasive internet of things issues. Academic Press, pp 183–210
42. Christian H, Agus MP, Suhartono D (2016) Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Comput Math Eng Appl 7(4):285–294



## Authors and Affiliations

**Gagandeep[1] · Jaskirat Kaur[2] · Sanket Mathur[1] · Sukhpreet Kaur[1] · Anand Nayyar[3] · Simar Preet Singh[4] · Sandeep Mathur[5]**

✉ Anand Nayyar
anandnayyar@duytan.edu.vn

Gagandeep
gaganpec@yahoo.com

Jaskirat Kaur
jaskiratkaur17@gmail.com

Sanket Mathur
rajeev.sanket@gmail.com

Sukhpreet Kaur
sukhpreet.4479@cgc.edu.in

Simar Preet Singh
dr.simarpreetsingh@gmail.com

Sandeep Mathur
sandeep2809@gmail.com

[1] Chandigarh Engineering College-CGC, Landran, Mohali, India

[2] Punjab Engineering College (Deemed to be University), Chandigarh, India

[3] Graduate School, Faculty of Information Technology, Duy Tan University, Da Nang, Viet Nam

[4] Bennett University, Greater Noida, India

[5] Noida International University, Greater Noida, India