ResumeScorePro
Enhancing Your Career Potential

EG/2020/4141 , Ranasinghe B.M.O.Y. EG/2020/4268 , Weerasingha W.M.N.S

***Abstract*— In the context of human resources and recruitment, this article discusses the development of ResumeScorePro, a machine learning solution for resume analysis, recommendations and rating. In order to standardize the recruitment procedure, the project uses machine learning algorithms, specifically Support Vector Machines (SVM) and Random Forest, to find an easy way to resume analysis and address the limitations of traditional resume screening methods. The ultimate objective is to increase the speed of application procedure and objectivity while developing a more transparent and balanced relationship between employers and job seekers.**

***Index Terms*—SVM – Support Vector Machine, HR – Human Resource**

I. Introduction

NORMAL resume screening procedure used in HR and recruiting is difficult, and vulnerable to interpretation. To resolve these issues, the project uses machine learning techniques for scoring and resume analysis. The need for more efficient hiring practices that result in more equitable hiring decisions is what makes this project so important. The usefulness of the selected algorithms, Random Forest and SVM, for resolving the sentiment analysis resume classification problem is examined. These algorithms' pros and cons are described.

Large companies receive many numbers of resumes for many types of job titles. So, HR managers do not have the ability to check every resume and find the best resume for a job. When the number of resumes is increasing, it will be very hard to manage the resume analyzing process. When checking a large number of resumes manually, the best resume can be missed. Because of that, an automatic resume selecting technique is very important for especially large companies.

The technique should choose the best resume. And also, that technique should have high precision. According to this machine learning project, the machine learning model could have the ability to select the most suitable resumes from the whole. When a resume is not satisfy the relevant qualifications for that particular job title, the model will predict it as zero. And also, if a resume is suitable for the job title, the model will predict it as one. Because of that, the HR manager should check only the resumes which are get one (1) from the machine learning model. It means, the HR manager can reject all the resumes which are get zero from the prediction of the machine learning model. For that, the model should be very precision one. If not, qualified candidates can be rejected without any manual checking.

.

The most significant problem that ResumeScorePro attempts to solve is the laborious and sometimes random process of resume testing, which can result in the rejection or mishandling of qualified candidates. So, the HR manager should select a suitable candidate for the suitable position. Our research aims to automate and standardize resume screening by applying machine learning, especially using Random Forest algorithms and SVM. This will introduce objectivity and specific requirements for a more rigorous review process.

The importance of ResumeScorePro is in its ability to optimize recruitment processes, resulting in enhanced efficiency and data-informed decision-making. The project objective is to close the gaps in traditional recruitment procedures by providing a more transparent and equitable experience for both companies and job seekers.
This initiative is important because it has the potential to improve the effectiveness of hiring practices and support decisions based on data. We will talk about the particular approach we used in this research, which includes Random Forest techniques and SVM.

II. Methodology

A. Data:

The open dataset used in our experiment was sourced from the website Kaggle. The dataset consists of 963 items, each of which has two columns: "category" and "resume." The score, which is either 1 or 0, denotes the standard of the resume and is represented by the 'category' column. The ' resume' column includes textual data that includes attributes like job experience, education, and talents. Our machine learning models are trained and tested using this dataset, which helps the algorithms identify patterns and correlations in the resumes. Here is the Kaggle data set link - https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset

B. Pre-processing:

A thorough pre-processing stage was carried out to ensure the quality and efficacy of our machine learning models. In the pre-processing step, mainly used techniques are data preprocessing and text preprocessing. In data pre-processing, duplicate and null-valued data are removed. And, in text preprocessing punctuation, uppercase, numbers, and special symbols are removed. To improve dataset integrity, duplicates were removed and potential biases were addressed by addressing null values. Text pre-processing included removing punctuation, changing the text's case, deleting stop words, digits, and special symbols, and standardizing word patterns using stemming techniques. To keep the dataset reliable, missing data management was done carefully.

From the whole data set, 80% of the data was used for the training step. The other 20% is used for the testing part. This methodical approach reduced the possibility of overfitting or underfitting and guaranteed a strong assessment of our models on unobserved data.

C. Algorithm:

SVM and Random Forest are two strong algorithms that are used for our machine learning project. These algorithms were selected based on how well they handled the categorization issue that arises from sentiment analysis of resumes. And, the problem of the project is a classification problem. So, those SVM and random forest algorithms were selected.

Support vector machines (SVM):

Finding the best decision boundaries is an area in which SVM performs as a reliable approach. To find the hyperplane that best divides several classes, it first transforms the input data into a higher-dimensional space. SVM works especially well in the context of ResumeScorePro to identify subtle patterns in resume sentiment, offering a dependable way to categorize resumes as positive (1) or negative (0). If the resume does not satisfy the suitable position, the machine learning model gives 0 (zero) and, if the resume satisfies the relevant qualifications, the model predicts the value as 1 (one).

Random Forest:

For greater accuracy and resilience, the Random Forest algorithm was used. The Random Forest algorithm is a combination learning method that is a combination of several decision trees and gives combined results of them. The adaptability and flexibility of Random Forest to manage a wide range of characteristics play a major role in the resume classification process's effectiveness. When the model predicted the value 1, the HR manager should only check those resumes. Then, the HR manager can find the best resume for the relevant position by checking a smaller number of resumes. This machine learning model is very useful for large companies.

D. Implementation:

The number of trees and maximum depth for Random Forest, as well as kernel types and regularization parameters for SVM, were taken into consideration when systematically exploring hyperparameters in the experiment conditions. These parameters were adjusted using cross-validation approaches, which balanced generalization and model complexity. To get a more accurate machine learning model, a tuning process would be applied.

After preprocessing, data set was converted text in to binary. The data set was unbalanced. It means, the data set got 1 and data set got zero is not balanced. For that, SMOTE was used to balance the data set. By increasing the size of the data set, the overfitting of the model can be removed.

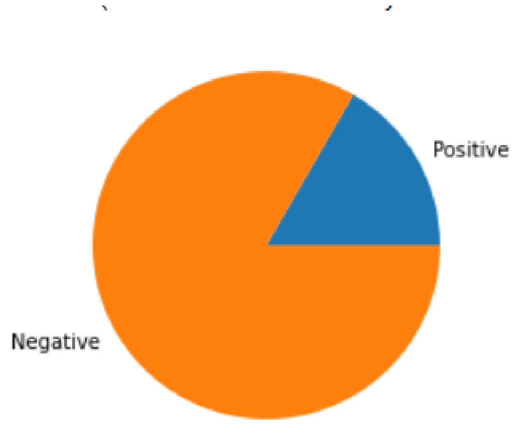

figure 1 : Overfitting of the data set

III. Results

- SVM Algorithm:

Training Scores:
Accuracy = 1.0
Precision = 1.0
Recall = 1.0
F1-score = 1.0
Test Scores:
Accuracy = 0.735
Precision = 0.733
Recall = 0.957
F1-score = 0.83

- Random Forest Algorithm:

Training Scores:
Accuracy = 1.0
Precision = 1.0
Recall = 1.0
F1-score = 1.0
Test Scores:
Accuracy = 0.735
Precision = 0.733
Recall = 0.957
F1-score = 0.83

the training score represents the models' ability to perfectly fit the training data, achieving ideal accuracy, precision, recall, and F1-score. The test score represents the models' ability to perfectly fit the test data of each algorithms.

- Models of Evaluation:

Problems with Overfitting:

For the training set, both SVM and random forest will get 100% accuracy. In practical accuracy cannot be 100%. So that is a weakness of this machine learning model. Because of that weakness, the model became overfitting.

Due to inadequate data, the Random Forest and SVM models both showed indications of overfitting. It's possible that the focus on outliers led to the development of an extremely complicated model with poor generalization.

- Tuning the Models :

Cross-validation:

To counteract overfitting, cross-validation techniques were used. With the SVM Algorithm, a mean score of 0.92740 was obtained by using an ideal cross-validation value of 9. With a mean score of 0.92833, the best cross-validation value for the Random Forest Algorithm was similarly 9.

Hyperparameter Tuning:

For the Random Forest and SVM algorithms, grid search cross-validation was carried out. For both techniques, a mean score of 0.7647058823529411 was obtained, suggesting a successful optimization.

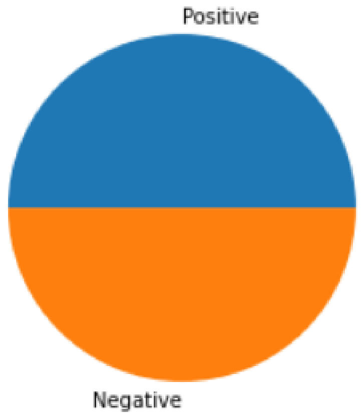

figure 2 : Balanced data set after the tuning

- Conclusions drawn from the Data

The model may be effectively utilized with both the Random Forest and SVM algorithms, exhibiting encouraging outcomes. In order to alleviate overfitting difficulties, future revisions will require a bigger and more balanced dataset. Performance can be improved by making changes to the support vector analysis, highlighting the requirement for a larger dataset. Depending on the performance of the model, HR managers can find the most qualified resume for each relevant job title.

As a conclusion, both data algorithms are suitable for the machine learning model. To avoid overfitting and for further modifications, a better data set can be used. Also, by adjusting the supporting vector analysis, better performances can be obtained from the machine learning model.

IV. Discussion

Ethical Considerations:

Fairness and biases: It's important to recognize the possibility of biases even when machine learning models seek to neutralize the hiring process. Although efforts were taken to guarantee the fairness of the training data, ongoing attention is required to avoid inadvertent biases in hiring choices.

Privacy: Privacy issues are brought up by the usage of personal information on resumes. Our initiative places a strong emphasis on the value of treating data ethically, making sure privacy laws are followed, and getting people's informed permission.

Discussion:

Algorithm Performance: The models' skill in resume sentiment analysis is shown by the obtained accuracy, precision, recall, and F1-score. But the problem of overfitting, which is a result of the small size of the dataset, highlights the need for a larger and more balanced dataset in order to improve model generalization.

Strategies for Optimization: Cross-validation and hyperparameter adjustment were used to effectively reduce overfitting, demonstrating the possibility of more gains with a larger dataset.

Algorithm Suitability: SVM and Random Forest both proved to be appropriate for the classification task, offering insightful information on the tone of resumes. Ongoing improvements are noted as possible areas for improvement, especially in support vector analysis.

V. Conclusion

To sum up, ResumeScorePro marks a substantial development in the area of automating and standardizing resume analysis for better recruiting procedures. Despite overfitting issues, the use of SVM and Random Forest algorithms demonstrates the possibility of revolutionary effects on the recruiting scene. When employing AI-based solutions in HR, ethical issues emphasize the dedication to justice, openness, and privacy. Acquiring a larger and more balanced dataset will be necessary for future developments to improve model training and generalization.