IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REVOLUTIONIZING TALENT ACQUISITION: ADVANCINGRESUME CLASSIFICATION WITH ITERATIVE LEARNING TECHNIQUE

[1]Asma Taj HA, [2]Amith KB, [3]Abdur Rehman

[1] Assistant Professor, [2] Student, [3] Student

Department of Information Science and Engineering,
Cambridge Institute of Technology, Bangalore, India

Abstract: With the increasing volume of digital resumes, efficient and accurate classification is essential for effective talent acquisition.This research delves into the innovative application of gradient boosting algorithms, a subset of machine learning techniques, for the intricate task of resume classification in the domain of talent acquisition. Gradient boosting methodologies, renowned for their adeptness in iteratively refining predictive models by combining weak learners, present a compelling avenue for bolstering the precision and efficiency of resume categorization systems.Moreover, the research endeavors to unravel the interpretability of gradient boosting models, shedding light on their role in fostering transparency and equity in the recruitment ecosystem. Through this multifaceted inquiry, this study not only advances the frontier of machine learning applications in talent acquisition but also underscores the transformative potential of gradient boosting in revolutionizing resume classification practices, thereby empowering organizations to make data-driven and equitable hiring decisions.

Index Terms **– Resume, Classification, Machine Learning.**

I. INTRODUCTION

In the contemporary landscape of talent acquisition, the process of resume classification stands as a pivotal component in identifying and attracting top-tier candidates. With the exponential growth of digital resumes, traditional manual screening approaches have become increasingly inefficient and prone to biases. Consequently, there arises a pressing need for innovative solutions that can streamline this process while ensuring fairness and accuracy.In response to this challenge, this project endeavors to explore the utilization of gradient boosting algorithms, a subset of machine learning techniques, for enhancing the efficacy of resume classification. By leveraging the iterative nature of gradient boosting, wherein weak learners are sequentially combined to create a robust predictive model, we aim to develop a system that can autonomously analyze and categorize resumes based on predefined criteria such as skills, experiences, and qualifications.Through this endeavor, we seek to not only optimize the efficiency of talent acquisition processes but also contribute to the advancement of transparent and equitable recruitment practices. This introduction sets the stage for a comprehensive exploration of the application of gradient boosting in resume classification, offering insights into its potential to revolutionize the recruitment landscape.

II. LITERATURE SURVEY

literature survey-Resume Classification using various Machine Learning Algorithms Machine Learning enables the path through which a computer can be trained to follow specific instructions again and again to make human life easy. The most common usage of machine learning is for the classification of objects[1]. In machine learning, iteration is important because models are exposed to new data and adapt accordingly.The Effect of Industry 4.0 and Artificial Intelligence on Human Resource Management. In today's market conditions, the importance of competition is obvious. Organizations must direct the right resources to the right investment to increase their competitive power and stay in the market. In this respect, the Human Resource Management (HRM) unit has also entered the digitalization phase. The digitalization phase in Human Resources (HR) has made significant progress, particularly in the recruitment process, with the help of Artificial Intelligence (AI) [2]. During this phase that creates a loss of value for the organization, searching for candidates among hundreds or even thousands of applications, selecting the most suitable one for the job, and placing the suitable ones in open positions within the institution;

III. METHODOLOGY

A. Proposed Method

This portion will elucidate the methodology and principles employed in constructing a classification model tailored for resume categorization, effectively matching candidates with suitable job profiles. The system operates through a series of distinct phases outlined below.

B. Data Gathering

Data Gathering includes collection of dataset from the data set provider website kaggle.com The datasets are classified and are structured datasets in which the data will be cleaned and stored in "UpdatedResumeDataSet .csv". 70% of the data is being used for training data and the remaining 30% will be used for test data.

C. Data Cleanup

The dataset contains a huge number of records that are very rough. Data cleaning will be done by removing any blank spaces, URLs, hashtags, special letters, and punctuation Stop words, are those words that don't play an important role in the sentence formation, such as "are," "we," "is," etc., are removed.

D. Label Encoding

In this step, each categorical label within the dataset, such as classes or categories, will undergo a transformation process known as label encoding. Label encoding involves converting categorical labels into numerical representations, which is essential for many machine learning algorithms to effectively process the data. By assigning a unique numerical value to each category or class, we enable the algorithm to interpret and analyze the data in a meaningful way.Label encoding begins by identifying the distinct categories or classes present in the dataset. Each category is then assigned a specific numerical value, typically starting from 0 and incrementing for each subsequent category. This numeric representation allows the algorithm to comprehend the relationship between different categories and perform calculations accordingly.Label encoding serves as a crucial preprocessing step in machine learning tasks, particularly in classification problems where categorical labels are involved. Once the labels are encoded, they can be seamlessly integrated into the training process of various machine learning models. Additionally, label encoding facilitates tasks such as model evaluation and result interpretation, as the algorithm can directly work with numerical representations of categorical labels.Various libraries and frameworks offer functionalities for label encoding, including scikit-learn's Label Encoder, providing flexibility and ease of implementation across different machine learning pipelines. By employing label encoding, we pave the way for subsequent stages of data processing and model training, contributing to the overall efficiency and effectiveness of the machine learning workflow.

E. TF-IDF Vectorization

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus).TF-IDF can be broken down into two parts TF (term frequency) and IDF (inverse document

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

frequency).Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency.

Inverse Document Frequency (IDF)

It is the measure of the importance of a word. Term frequency (TF) does not consider the importance of words. Some words such as' of', 'and', etc. can be most frequently present but are of little significance. IDF provides weightage to each word based on its frequency in the corpus D.

$$IDF(w, D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

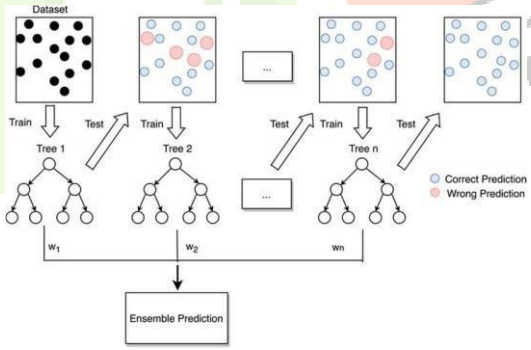
This process identifies and highlights key words and phrases associated with employment roles and skill sets by analyzing their frequency within the text. These highlighted terms, characterized by their highest occurrence rates, can then be utilized to extract crucial vocabulary for training datasets across different machine learning algorithm.

TF-IDF Vectorization = TF × IDF

This process involves identifying and highlighting the most frequent words related to job roles and skill sets within a dataset. These highlighted words serve as key indicators of the essential attributes and qualifications sought after in candidates. By extracting these high-frequency terms, we can create a curated list of significant words that can be utilized to train machine learning algorithm effectively. This curated list forms the basis for building classification model capable of accurately categorizing and analyzing job-related data, facilitating better decision-making in talent acquisition and recruitment processes.

F. Applying Classification Algorithm To Dataset

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize



this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

Fig.1: Shows the Architecture of the Gradient Boosting classifier

IV. METHODOLOGY FLOWCHART

Understanding the methodology is crucial, so we've created a flowchart to provide a visual representation of how the system operates. Figure 2 illustrates the step-by- step process of the system, making it easier to comprehend the overall flow and sequence of activities involved.

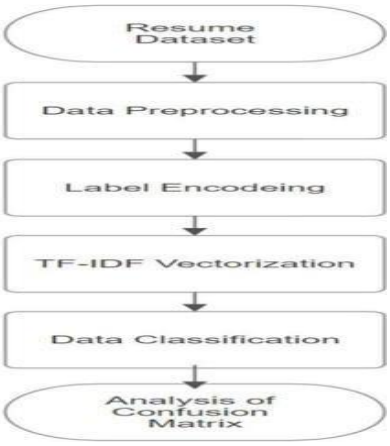


Fig.2: Methodology Flowchart

V. RESULTS

The source of the dataset is kaggle.com. The dataset underwent preprocessing, which involved label encoding and stop word removal. Fig. 3 displays the data before preprocessing. Fig. 4 displays the data after preprocessing

Fig.3: Displays The Data Before Preprocessing.

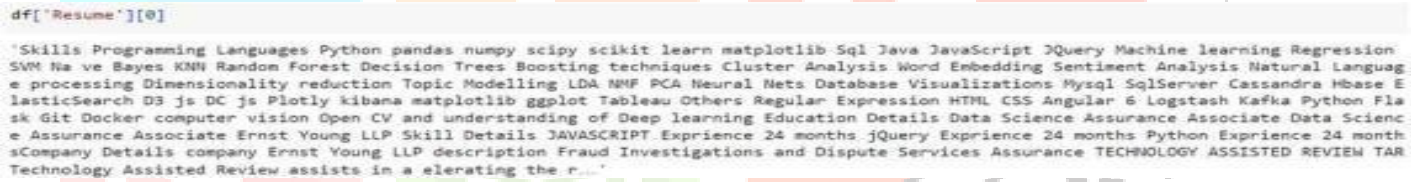


Fig.4: Displays The Data After Preprocessing.

Following data preprocessing, the dataset undergoes TF-IDF transformation to assign importance scores to each word, enabling the arrangement of the word matrix based on ascending order of significance. In the resulting TF-IDF vectorization output, columns representing words with a term frequency exceeding 5000 are depicted, as illustrated in Figure 5.3 with output like 'skills': 6080, 'programming': 5160, 'languages': 3713, 'python': 5266, 'pandas': 4742, 'numpy': 4490, 'scipy': 5857,'scikit': 5855, 'learn': 3756, 'matplotlib': 4069, 'sql': 6239, 'java':3537, 'javascript': 3539, 'jquery': 3581...

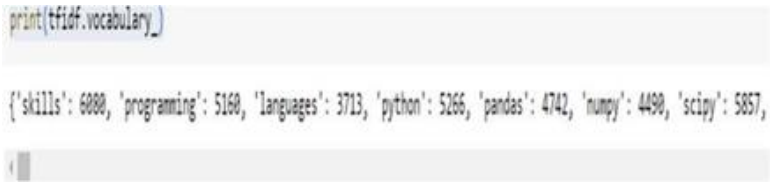


Fig.5: Output of TD-IDF Vectorization.

Figure 5 illustrates the primary 25 job categories under evaluation, comprising
.['Data Science', 'HR', 'Advocate', 'Arts', 'Web Designing', 'Mechanical Engineer', 'Sales', 'Health and fitness', 'Civil Engineer', 'Java Developer', 'Business Analyst','SAP Developer', 'Automation Testing', 'Electrical Engineering','Operations Manager', 'Python Developer', 'DevOps Engineer','Network Security Engineer', 'PMO', 'Database', 'Hadoop', 'ETL Developer', 'DotNet Developer', 'Blockchain','Testing'], dtype=object)

After data preprocessing and organization, data classification into diverse job roles and skill clusters can be accomplished utilizing the gradient boosting algorithm. The resultant systems exhibit remarkable accuracy in predicting true values when analyzing the confusion matrix. The subsequent visual representations illustrate the confusion matrices for various classifications, highlighting true positive instances where the model accurately predicts job profiles as expected values. Figure 6 specifically displays the confusion matrix encompassing all job categories.

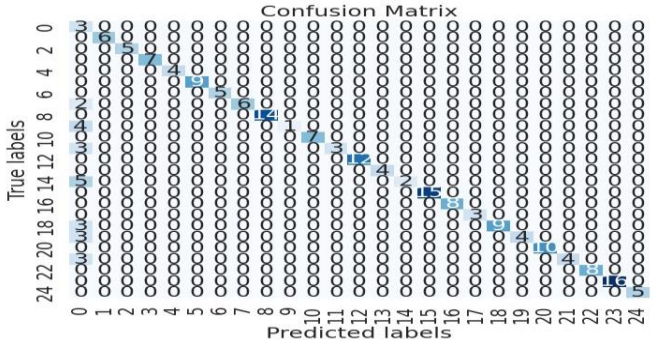


Fig.6: confusion matrix encompassing all job categories

Accuracy	
Training	90.377
Testing	88.083

VI. CONCLUSION AND FUTURE WORK

Analysis model using the Gradient Boosting Classifier have shown promising results. By leveraging ensemble learning techniques and combining the strengths of multiple weak learners, we were able to build a robust model capable of accurately classifying the sentiment of text data.Throughout the project, we explored various preprocessing techniques, feature engineering methods, and model hyperparameter tuning to optimize the performance of the Gradient Boosting Classifier. We observed significant improvements in both accuracy and generalization capability, demonstrating the effectiveness of our approach in handling sentiment analysis tasks.While the current sentiment analysis model has achieved satisfactory performance, there are several avenues for future research and improvement:Exploration of Deep Learning Models: Investigate the applicability of deep learning architectures, such as recurrent neural networks (RNNs) or transformers, for sentiment analysis tasks. These models have shown promising results in capturing complex linguistic patterns and context dependencies.Real-time Deployment and Scalability: Optimize the model for real-time deployment and scalability by leveraging cloud- based platforms and containerization technologies. This would enable seamless integration into production environments and support high-throughput inference for large-scale applications.Continuous.Model Monitoring and Evaluation: Implement a robust monitoring system to continuously evaluate the performance of the sentiment analysis model in real-world scenarios. Regular model retraining and updates can ensure its effectiveness and adaptability to evolving language trends and user preferences.

REFERENCES

[1] Mohamed A, Bagawathinathan W, Iqbal U, Shamrath S, Jayakody A (2018) Smart talents recruiter-resume ranking and recommendation system. In: 2018 IEEE international conference on information and automation forsustainability (ICIAfS). international conference on nascent technologies in engineering (ICNTE).

[2] S. M, I. P. B, M. Kuppala, V. S. Karpe and D. Dharavath, "Automated Resume Classification System Using Ensemble Learning," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 1782-1785,

[3] B. Gunaseelan, S. Mandal and V. Rajagopalan, "Automatic Extraction of Segments fromResumes using Machine Learning," 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342596.

[4] Gaur, B., Saluja, G. S., Sivakumar, H. B., & Singh, S. (2020). Semi-supervised deep learning based named entity recognition model to pars education section of resumes.

https://www.semanticscholar.org/paper/Semi-supervised-deep- learning-based-named-entity-to-Gaur-Saluja/eeaf25e7498b9db244404791c51e16134e981d7f

[5] Li, C., Fisher, E., Thomas, R., Pittard, S., Hertzberg, V., & Choi,J. D. (2020). Competence-Level prediction and Resume- Job_Descriptionmatching using Context-Aware transformer models.https://www.semanticscholar.org/paper/Competence- Level-Prediction-and-Matching-Using-Li-Fisher/5977016b91f3dafa72ce24d4d343ddbee91a285e

[6] Humanverification. (n.d.). https://www.semanticscholar.org/paper/RecruitPro%3A-A-Pretrained-Language-Model-with-Prompt-Fang- Qin/c7d65ad006341d57be3080dd6572c307779c2c4b

[7] Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., & Viviani, M. (2017). WoLMIS: alabor market intelligence system for classifying web job vacancies. https://www.semanticscholar.org/paper/WoLMIS%3A-a-labor- market-intelligence-system-for-web-Boselli-Cesarini/ac77e18f90d8566625e051bee4e0b4df818771e3

[8] Humanverification. (n.d.-b). https://www.semanticscholar.org/paper/Automatic-Software- Engineering-Position-Resume-Word-Pant-Pokhrel/6a2d4eccd8c0679afa3f9c83725ee5e84dbd378f

[9] Humanverification. (n.d.-c). https://www.semanticscholar.org/paper/Learning- Representations-for-Soft-Skill-Matching-Sayfullina- Malmi/52a9dd30825039a98e720dfd1c443c104e1f156dK.

[10] Sinha, A. K., Akhtar, M. a. K., & Kumar, M. (2023). Automated Resume Parsing and Job Domain Prediction using Machine Learning. Indian Journal of Science and Technology, 16(26),1967– 1974. https://doi.org/10.17485/ijst/v16i26.880

[11] A. Zaroor, M. Maree and M. Sabha, "JRC: A Job Post and Resume Classification System for Online Recruitment," 2017IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, USA, 2017, pp. 780-787, doi: 10.1109/ICTAI.2017.00123.

[12] Maitri Dipak Amin, Aishwarya Chandrashekhar Harkare, Nitya Singh Parmar, Ruchika Rajesh Wadhwa, Rajesh A. Patil, "Real Time Data based Automated Resume Classification and Job.Matching using SVC, Jaccard Index and Cosine Similarity", 2023 7th International Conference on Computer Applications in Electrical Engineering-Recent Advances (CERA), pp.1-6, 2023.

[13] Sonali Mhatre, Bhawana Dakhare, Vaibhav Ankolekar, Neha Chogale, Rutuja Navghane, Pooja Gotarne, "Resume Screening and Ranking using Convolutional Neural Network", 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), pp.412-419, 2023.

[14] Raj Pandey, Divya Chaudhari, Sahil Bhawani, Omkar Pawar, Sunita Barve, "Interview Bot with Automatic Question Generation and Answer Evaluation", 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), vol.1, pp.1279-1286, 2023.

[15] Muskan Sharma, Gargi Choudhary, Seba Susan, "Resume Classification using Elite Bag-of-Words Approach", 2023 5th International Conference on Smart Systems and InventiveTechnology (ICSSIT), pp.1409-1413, 2023.

[16] O Pandithurai, D Jayashree, D K Aarthy, R Jaishree, K Bhavani, T Dharani, "Smart Job Recruitment Automation Using Location Based Filtering", 2021 International Conference on Advancementsin Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp.1-4, 2021.

[17] Sujit Amin, Nikita Jayakar, Sonia Sunny, Pheba Babu, M. Kiruthika, Ambarish Gurjar, "Web Application for Screening Resume", 2019 International Conference on Nascent Technologies in Engineering (ICNTE), pp.1-7, 2019.