RESUME SCREENING USING MACHINE LEARNING

MUNGI NAGA VENKATA SAI RAGHAVENDRA[1]
PG student, Department of Computer Science and System Engineering
Andhra University
Visakhapatnam, India

Abstract— Resume screening is the process of analyzing the resumes where the candidates apply for the different types of jobs where the company feel the tedious job to find the appropriate candidate due to the complexity in resumes formats since it has different styles. As a result, selecting applicants for the appropriate job within a company is a difficult task for recruiters. We can extract the key information from the CV using NLTK, Natural Language Processing (NLP) techniques to save time and effort. This system could work with a large number of resumes for classifying the right categories using different classifiers like KNN, SVM, MLP, LR. Furthermore, this system attempts to find the accuracy and performance of the proposed methodology and incorporate it in the IT firms and other regulations for the prevention of manual screening and establish a safe allocation of resources for the companies.

Index Terms—Resume, CV, NLTK, NLP, KNN, SVM, MLP, LR

1. Introduction

Hiring the right talent is a challenge for all businesses. This challenge is magnified by the high volume of applicants if the business is labor-intensive, growing, and facing high attrition rates. An example of such a business is that IT departments are short of growing markets. In a typical service organization, professionals with a variety of technical skills and business domain expertise are hired and assigned to projects to resolve customer issues. This task of selecting the best talent among many is known as Resume Screening. Typically, large companies do not have enough time to open each CV, so they use machine learning algorithms for the Resume Screening task and by this unemployment rate [8] also reduced with efficient hiring. Machine learning is a field in which we train a model with data to anticipate the intended outcome when new data is submitted. Natural language processing (NLP) is a commonly used to screen resumes. Natural language refers to how humans communicate with one another. In the NLP the system enables us to find the text based on the English dictionary in the same way as humans. NLP combines statistical, machine learning, and deep learning models of human language with computational linguistics-based rule-based modeling, here we need to check for the data from different formats which are either in the form of the document or either in the form of the audio data and understanding the whole meaning of it. The number of applications is in the millions, making it a time-consuming chore to sort through them. Here we need a machine learning algorithm that can give a better way of understanding and also can full fill the requirements according to the requirement in the industry. The proposed system takes a CSV file as input which contains different categories and resumes based on the category and features of the resume the accuracy and performance are calculated using different machine learning classifiers.

2. Literature Survey

A. Machine Learning approach for automation of Resume Recommendation System

Choosing the best candidates from the pool here to perform these types of tasks different NLP techniques such as bigram trigram and n gram and text classification are used, this model used Machine Learning to perform the classification using the algorithm [1].

B. Skill Finder: Automated Job-Resume Matching System

API for web services [9]. This information is then utilized to score the students' resumes based on the skills required for the job, using Named Entity Recognition (NER) software such as Apache OpenNLP [10] and Stanford Name Entity Recognizer [11].

C. Resume NET: A Learning-based Framework for Automatic Resume

Quality Assessment

Yong Luo [10] produced a custom dataset of out of is categorized in two categories: positive and negative, with 33 and 89 resumes identified as positive and negative, respectively [3].

D. Web Application for Screening Resume

The goal was to create a web application for resume screening using 220 resumes, 200 of which were utilized for training and 20 for testing, and the web application was separated into three sections.

- Job Applicant side
- Server-Side
- Recruiter Side

The applicant will supply his or her résumé on the applicant side, which will be processed on the server side and then trained using the NLP Pipeline, which uses SpaCy, an NLP framework [6]. On the recruiter's side, the resume rank list will be displayed, which was determined using a score calculator,so that the recruiter may choose the best candidate for the job.

E. Design and Development of Machine Learning based Resume Ranking System

The system proposes a technique in which the candidate submits his or her resume following an interview here the face-based technique was used. After the resume is submitted, NLP techniques are used to extract the necessary abilities from the resume, then TF- IDF vectorization is used to transform the words into vectors so the machine can interpret them. The KNN algorithm [5] is used to identify the resume that most closely fits the JD provided by the recruiter. The system has a parsing accuracy of 85 percent on average [4].

F. Differential Hiring using Combination of NER and Word Embedding

The NER model is used to extract useful entities from documents, which is enhanced by the word2vec model by making the system more generic and the similarity is calculated using the cosine similarity algorithm [7].

G. Al-Otaibi et al., [12] provided a detailed survey of job recommendation service. They discussed the steps involved in the recruiting process used by any organization. How the e-recruitment portal is helping to the organization, what factor of the candidate may lead to getting selected and many other relevant recruitment processes are explained.

3. Implementation Study

Resume screening is a strategy largely used by Big Tech businesses to sort through a huge number of resumes, rank them according to resume strength or relevance to the job description, and then filter them. The student seeking for the position, on the other hand, has no understanding why his resume was turned down or how he might modify his CV to make it more relevant and remarkable. There is no technology available right now that would benefit students and help them create their resume.

To resolve the given issue statement, the machine learning model will be employed. It will read the student's resume and extract information such as abilities and credentials. It also takes connections to the student's GitHub and LinkedIn profiles for more information, from which it can extract the student's contributions in a variety of fields. The student must also state which job role he or she is applying for. The model is trained using a set of job descriptions and skill sets.

3.1 Data Pre-Processing:

This process would shortlist the resumes provided as input to eliminate any special or garbage characters from the resumes. During the cleaning process, all unique characters, digits, and single-letter words are removed. We had a clean dataset after these operations, with no unique characters, digits, or single- letter words. To tokenize the dataset, NLTK tokenizers are employed.

3.2 Proposed Methodology:

The model's concept will be trained with existing data gathered from the Kaggle open platform. The first model, either K-Nearest Neighbor, SVM, MLP and LR. will help us predict what kind of job role our resume is best suited for, while the second model, cosine similarity, will check the user's input of what job role they want, and the recommendation system will provide it based on that. The following is the control flow: At the front

end, the candidate uploads their resume; the resume is then passed to the resume parser, which is a pipeline of NLP algorithms that extracts important information. information from the resume; and finally, adding more value to the overall extracted data from vectors and providing it to the Machine learning Model for tagging.

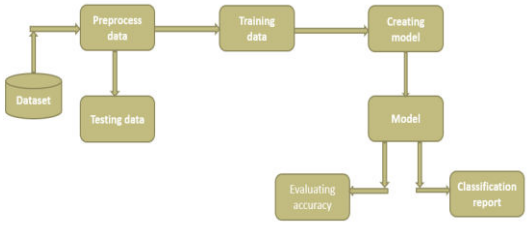


Fig 1: - Architecture Diagram

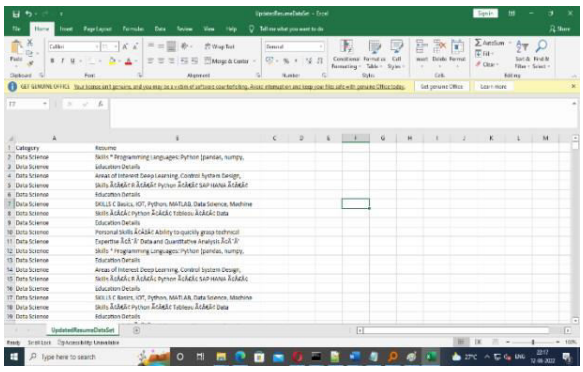


Fig 2: - Dataset Description

4. Algorithms Used

4.1 KNN Algorithm

The K Nearest Neighbor algorithm [13] is a non-parametric algorithm, which means it makes no data assumptions. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and uses it to categories it later. During the training phase, the KNN algorithm simply saves the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

Algorithm: The Traditional KNN Algorithm

Input: the training set D, test object x, category label set C
Output: the category c_x of test object x, c_x belongs to the C

```
1  begin
2    for each y belongs to D do
3      calculate the distance D(y, x) between y and x
4    end for
5    select the subset N from the data set D,
     the N contains k training samples which are the k
     nearest neighbors of the test sample x
6    calculate the category of x:
     c_x = arg max_{c∈C} Σ_{y∈N} I(c = class(y))
7  end
```

Fig 3: - KNN Algorithm Step by Step Process

4.2 SVM Algorithm

A supervised machine learning approach called Support Vector Machine (SVM) [14] is used for both classification and regression. Although we also refer to regression problems, classification is the most appropriate term. Finding a hyperplane in an N-dimensional space that clearly classifies the data points is the goal of the SVM method. The number of features determines the hyperplane's size. The hyperplane is essentially a line if there are just two input features. The hyperplane turns into a 2-D plane if there are three input features. Imagining something with more than three features gets challenging.

Algorithm 1 Training an SVM

Require: X and y loaded with training labeled data, $\alpha \Leftarrow 0$ or $\alpha \Leftarrow$ partially trained SVM

1: $C \Leftarrow$ some value (10 for example)
2: **repeat**
3: **for all** $\{x_i, y_i\}, \{x_j, y_j\}$ **do**
4: Optimize α_i and α_j
5: **end for**
6: **until** no changes in α or other resource constraint criteria met

Ensure: Retain only the support vectors ($\alpha_i > 0$)

Fig 4: - SVM Algorithm Step by Step Process

4.3 Logistic Regression

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable has a binary solution. Similar to all other types of regression systems, Logistic Regression is also a type of predictive regression system. Logistic regression is used to evaluate the relationship between one dependent binary variable and one or more independent variables. It gives discrete outputs ranging between 0 and 1.

4.4 MLP Algorithm

Multi-Layer Perceptron(MLP) is the simplest type of artificial neural network. It is a combination of multiple perceptron models. Perceptrons are inspired by the human brain and try to simulate its functionality to solve problems. In MLP, these perceptrons are highly interconnected and parallel in nature. This parallelization helpful in faster computation.An MLP is characterized by several

layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network.

1: choose an initial weight vector ~w
2: initialize minimization approach
3: **while** error did not converge **do**
4: **for all** (~x, ~d) ∈ D **do**
5: apply ~x to network and calculate the network output
6: calculate ∂e(~x)
7: **end for**
8: calculate ∂E(D)
9 for all weights summing over all training patterns
10 perform one update step of the minimization approach
11: **end while**

Fig 5: - MLP Algorithm Process

5. Results and Evaluation Metrics

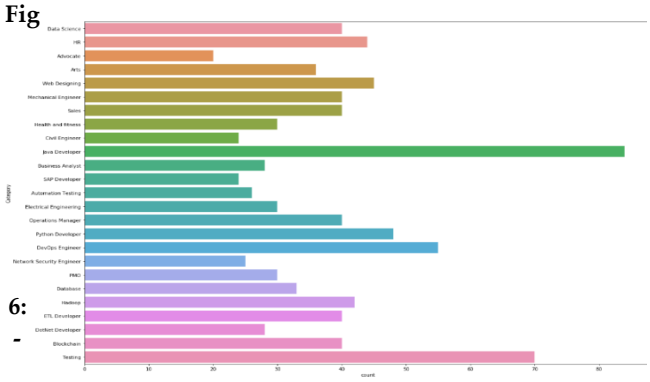


Fig 6: - Dataset Category Values and Category Count Representation In Bar Chart

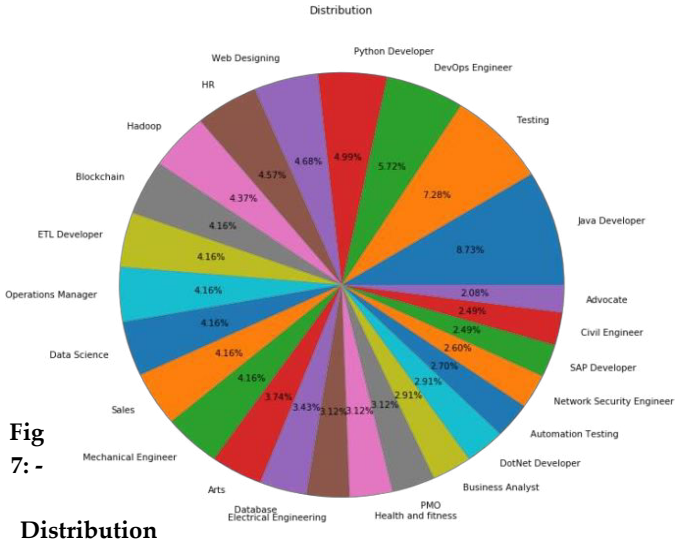


Fig 7: - Distribution Graph

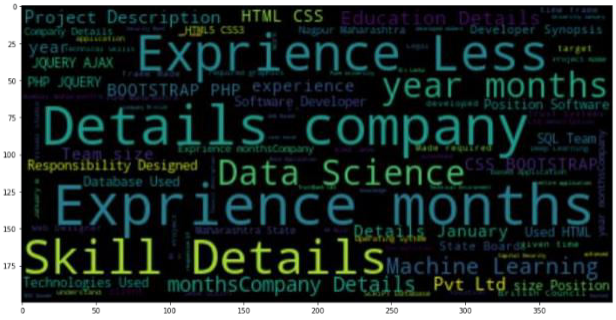


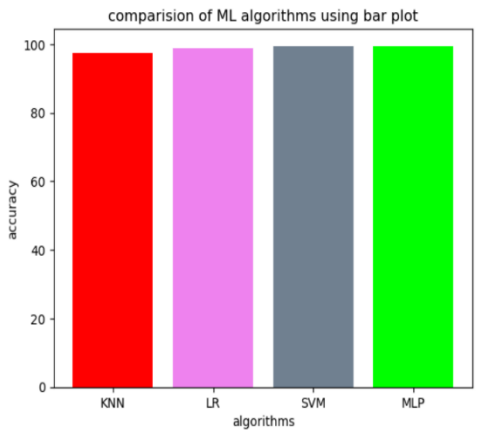
Fig 8: - Word Cloud Graph

	Category	Resume	clean text
0	6	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	6	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	6	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	6	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Table...	Skills R Python SAP HANA Table...
4	6	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...
...	...	...	...
957	23	Computer Skills: â□¢ Proficient in MS office (...	Computer Skills Proficient in MS office Wo...
958	23	â□□ Willingness to accept the challenges. â□□ ...	Willingness to a ept the challenges Po...
959	23	PERSONAL SKILLS â□¢ Quick learner, â□¢ Eagerne...	PERSONAL SKILLS Quick learner Eagernes...
960	23	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...	COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power Po...
961	23	Skill Set OS Windows XP/7/8/8.1/10 Database MY...	Skill Set OS Windows XP 7 8 8 1 10 Database MY...

Fig 9: - Cleaned data from the dataset

	model	accuracy
0	KNN_acc	97.4093
1	LR_acc	98.9637
2	SVM_acc	99.4819
3	MLP_acc	99.4819

Fig 10: - Accuracy of ML Classifiers



	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	3
2	1.00	1.00	1.00	5
3	1.00	1.00	1.00	9
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	5
6	1.00	1.00	1.00	9
7	1.00	1.00	1.00	7
8	1.00	0.91	0.95	11
9	1.00	1.00	1.00	9
10	0.89	1.00	0.94	8
11	1.00	1.00	1.00	9
12	1.00	1.00	1.00	5
13	1.00	1.00	1.00	9
14	1.00	1.00	1.00	7
15	1.00	1.00	1.00	19
16	1.00	1.00	1.00	3
17	1.00	1.00	1.00	4
18	1.00	1.00	1.00	5
19	1.00	1.00	1.00	6
20	1.00	1.00	1.00	11
21	1.00	1.00	1.00	4
22	1.00	1.00	1.00	13
23	1.00	1.00	1.00	15
24	1.00	1.00	1.00	8
accuracy			0.99	193
macro avg	1.00	1.00	1.00	193
weighted avg	1.00	0.99	0.99	193

Fig 11: - Comparison of ML Algorithms Using Bar Plot

Fig 12: - MLP Algorithm Classification Report

6. Conclusion

This paper examines a variety of machine learning model such as KNN, SVM, logistic regression and MLP, to detect, identify, and categories diverse resumes. And here we achieve the better accuracy and we implement a web interface to screen the resumes and analyses the type of job related to resume, MLP outperforms other approaches like KNN, SVM, Logistic Regression.

7. References

[1] Pradeep Kumar Roy, Vellore Institute of Technology, 2019. A Machine learning approach for automation of resume recommendation system, ICCIDS 2019. 10.1016/j.procs.2020.03.284.

[2] Thimma Reddy Kalva, Utah State University, 2013. Skill-Finder: Automated Job-Resume.

[3] Based Framework for automatic resume quality Suhjit Amin, Fr.Conceicao Rodrigues Institute of Technology, 2019. Web Application for Screening resume, IEEE DOI: 10.1109/ICNTE44896.2019.8945869.

[4] Ashwini K, Umadevi V, Shashank M Kadiwal,Revanna, Design and Development of e Learning based Resume Ranking.

[5] Riza tana Fareed, rajah V, and Sharadadevi kaganumat "Resume Classification and Ranking using KNN and Cosine Similarity" In 2021 International Journal of Engineering.

[6] 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), DOI: 10.1109/ICNTE44896.2019.8945869.

[7] Suhas H E, Manjunath AE, "Differential Hiring using Combination of NER and Word Embedding", In 2020 International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Vol.9

[8] Centre for Monitoring Indian Economy Pvt Ltd. (CMIE),2022. The unemployment rate in India.

[9] Howard, J.L., Ferris, G.R., 1996. The employment interview context: Social and situational influences on interviewer decisions Xavier Schmitt, Sylvain Kubler, Jer my Robert, Mike Papadakis, Yves LeTraon University of Luxembourg, Luxembourg Replicable Comparison Study NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate.

[10] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[11] Mikheev, Andrei; Moens, Marc; Glover, 1999. "Named Entity Recognition without Gazetteers." Proceedings of EACL '99. HCRCLanguage Technology Group, University of Edinburgh, http://acl.ldc.upenn.edu/E/E99/E99-1001.pdf.

[12] Al-Otaibi, S.T., Ykhlef, M., 2012. A survey of job recommender systems. International Journal of Physical Sciences 7, 5127–5142.

[13] Bhushan Kinge*1, Shrinivas Mandhare2, Pranali Chavan3, S. M. Chaware4 , Resume Screening Using Machine Learning and NLP : A

Proposed System, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, ISSN : 2456-3307 UGC Journal No : 64718.

[14] Scholkopf, B., Smola, A.J., Bach, F., et al., 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press.