Machine Learned Resume Scrutiny for Job Recommendations

[1]Sahil Sanghavi, [2]Usaid Shaikh, [3]Chetan Daulani, [4]Prof. Ankita Karale

[1]Student, [2]Student, [3]Student, [4]Professor
[1]Computer Engineering, [2]Computer Engineering, [3]Computer Engineering, [4]Computer Engineering,
[1]SITRC, Nasik, India, [2]SITRC, Nasik, India, [3]SITRC, Nasik, India, [4]SITRC, Nasik, India.

Abstract: This study adopts machine learning and text mining technology to analyse a candidate profile and resume, developing a system that can be applied where numerous applicants seek to match with the job vacancies provided by companies. The developed system conducts personal competitiveness analysis, personality analysis, and gives recommendations according to the resumes applicants submit.

Keywords **- natural language processing, artificial intelligence, text mining, recommendation system, data analysis.**

I. INTRODUCTION

Job seekers who want to develop in their careers are excited about looking for new chances. The issue is that the current method isn't very adaptable, efficient, or time-saving. Our technology helps candidates save time by allowing them to post their resume in whichever format they desire. All of the information in the CV will be detected by our system from the individual's social profile, resulting in the best applicant for that particular position. The candidate will be delighted as well since he will be hired by a firm that values his skills and abilities. On the other hand, we provide the client organisation the same level of adaptability.

Candidates in the current system are dissatisfied with the work that the current system chooses based on their qualifications. Let me explain what a 5:1 ratio means: if 5 employees are hired, then one of them can be content with his or her work. As an example, consider the following: If I am a decent Python developer and a corporation hires me and forces me to focus on Java, my Python expertise would be rendered worthless. On the other hand, if there is a vacant position in a company, the owner of the company would choose the best potential choice for that vacancy. As a result, our device would serve as a handshake between the two organisations. The company that chooses the best candidate possible and the candidate that prefers the best position possible based on his or her talents and abilities.

According to a survey report conducted by Glassdoor in 2015, approximately 80% of millennial job applicants check whether or not the culture of the company they intend to apply to suits them before even evaluating the development potential of the company.[2]

II. LITERATURE SURVEY

During graduation season, or the traditional work transition time between the end of one year and the start of the next, job seekers eager to progress to the next level of their careers are on the lookout for opportunities. Job seekers can face a variety of circumstances during the job search process: 1. they are unable to thoroughly review the content and requirements of each job vacancy due to the large number of job vacancies available at the time; 2. they are unaware of their own conditions and competencies; also 3. Their alignment (based on various indicators such as experience, qualifications, and personal characteristics) with the criteria and essence of the work position they apply for determines whether or not they are accepted to the next-stage interview after undergoing the initial on-site interview.

According to a Glass door poll taken in 2015, about 80% of millennial job seekers examine if the company's culture fits them prior to actually analysing the company's development prospects [2]. This makes a strong case of comprehending the stated recognition as well as the organizational values. The aforementioned situation is also reflected in companies' talent recruitment process. HR employees must spend a significant amount of time and effort examining resumes during the height of the resume submission process, and are consequently prone to making rash decisions.

To summarise, firms have a constant difficulty in identifying appropriate individuals from a sea of candidates and inviting them to interview. Repetitive jobs have become a focus of artificial intelligence development as the technology has become more prevalent. The system will be used as a reference for job matching in order to improve its success rate, which will benefit both sides.

i. Recommendation systems

A personalised suggestion feature is standard in most recommendation systems. A research [5] even claimed the possibility of developing a tailored position suggestion list based on job seekers' specific preferences, highlighting the efficiency of the customised function in job searching.

ii. Textual analysis

Textual analysis, in terms of natural language processing, is a complex approach for analysing papers in any discipline, whether it is used for subject prediction, paper grouping, or keyword extraction.

iii. Data analysis

The realisation of customised data has become a trend in the present setting of widespread use of big data analysis. Based on current or historical data, our analysis offers predictions regarding future events. Forecasting is nothing more than a guess. Its precision is determined by how much comprehensive information you have and how far you go into it.

III. MODELS AND TECHNOLOGIES

i. Natural Language Processing (NLP)

We use the term Natural Language Processing — or NLP for short — to refer to any form of computer-assisted natural language manipulation. NLP is majorly a subset of AI and Machine Learned linguistics. At one extreme, comparing various writing types may be as easy as counting word frequencies. NLP, on the other hand, entails "understanding" full human utterances, at least to the point of being able to respond to them in a useful manner.

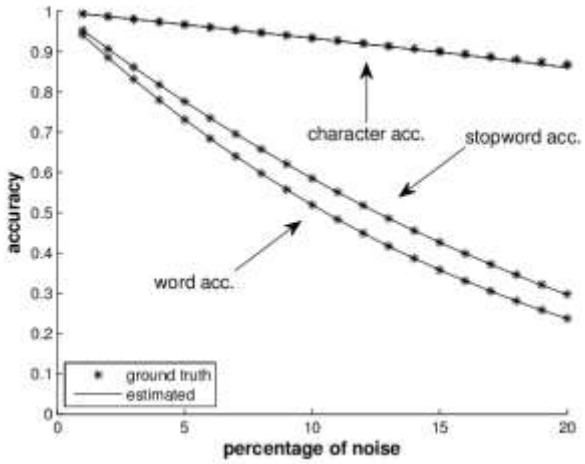

Fig. 1. *Using NLP - Morphological, Syntactic and Semantics to improve recognition accuracy [16]*

By merely including more input data, systems dependent on automatically learning the rules can be made more reliable. Handwritten rule-based programmes, on the other hand, can only be made more effective by increasing the sophistication of the rules, which is a much more challenging challenge. There is, in fact, a limit to the sophistication of systems built on handwritten laws, after which the systems become increasingly unmanageable. Creating additional data to feed into machine-learning programmes, on the other hand, actually necessitates an improvement in the number of man-hours employed, with no noticeable rise in the sophistication of the annotation method.

NLP requires the following constraints for parsing:

1. Morphological Analysis

In linguistics, morphology is the analysis and explanation of how words are produced in natural language. In this phase, the sentence is broken down into tokens—the smallest unit of words—in this step, which defines the basic form of the expression. For example, unusually is made up of the prefix un-, the stem regular, and the suffix affably.

The standard method is to divide a complex system into components, separate the parts whose inputs are crucial to the success (dropping the 'trivial' elements), and solve the simpler system for desired scenarios. The downside of this approach is that certain real-world phenomena lack apparent trivial elements and thus cannot be generalised. Without a simplification stage, morphological analysis operates backward from the output to the function internals. The research thoroughly accounts for the system's experiences.

2. Syntactic Analysis

The aim of syntactic analysis is to determine the sentence's syntactic structure. It is also known as Hierarchical analysis/Parsing and is used to understand a sentence, assign token classes to grammatical phrases, and assign a syntactic form to it. The aim of this step is to extract exact meaning, or dictionary meaning, from the text. Syntax review examines the document for context by comparing it to formal grammar codes. For example, a semantic analyzer will reject the sentence "hot ice cream". In this context, syntactic analysis or parsing can be described as the method of analysing strings of symbols in natural language in accordance with formal grammar rules. The term "parsing" derives from the Latin word "pars," which means "section."

3. Semantic Analysis

The evolution of representation patterns for the meaning of language inputs is the focus of semantic analysis. It explains how to determine a sentence's structure from the content of its elements. It generates a logical query, which is sent into the Sql Query Builder. For user tokens and user input symbols, it is another sort of semantic word representation. The goal of semantic analysis is to derive the text's particular definition, or dictionary interpretation. The semantic analyzer's purpose is to look for context in the text.

As a result, semantic interpretation can be separated into two parts:

a. Investigating the meanings of particular words

It is the first stage of semantic research in which the meaning of particular words is investigated. This is referred to as lexical semantics.

b. Investigating the interaction between individual terms

The individual words would be merged to have context in sentences in the second chapter.

ii. The DiSC Model

The DiSC model, which was developed in the 1920s by psychologist William Moulton Marston, is a common, concise, systematic, and reasonably simple approach to evaluate behavioural styles and interests.[1]

The method categorises people's actions into four categories (Dominance, Influence, Stability, and Conscientiousness) based on their interests on two scales:

- Task versus People.
- Fast-Paced versus Moderate-Paced.

Type	**Their Behaviour**
Strong Dominance	Idiopathic, Listening is more important than talking. Opinionated, obstinate, forceful, and determined
String Influence	More discussion is needed, Emotional, persuasional, political, Persuasive and animated
Strong Steadiness	Rather than telling, ask questions. Consistent, consultative, patient, averse to change, and reserved
Strong Conscientiousness	Understand the law. Careful, thoughtful, rigorous, tactful form

Table 1. *The Four Quadrants of the DISC Model*

iii. Scikit Learn

Scikit learn seeks to provide easy and effective learning solutions that are open to all and reusable in a variety of contexts: machine learning as a flexible platform for science and engineering. A learning problem, in general, considers a series of n samples of data and attempts to predict properties of unknown data. If a dataset contains more than one number, such as a multidimensional entry (also known as multivariate data), it is said to have multiple attributes or characteristics. We will use supervised learning via Scikit Learn in this project. Supervised learning occurs where the data contains additional properties that we want to forecast.

This issue may be one of two things:

1. ***classification*****:** samples belong to two or three groups, and we want to learn how to predict the class of unlabeled data from previously classified data. The digit recognition problem is an example of a classification problem in which the aim is to assign each input vector to one of a finite number of discrete categories.
2. ***regression*****:** If the intended output is made up of one or more continuous variables, the task is referred to as regression. A regression issue will be the estimation of a salmon's length as a function of its age and weight.

Unsupervised learning is a type of deep learning in which the instructional data is made up of a sequence of feature vector xi that do not have matching goal quantities. The intent of such studies will be to locate similarity between two occurrences within the input (knn) or to estimate the spread of data within the training data (volume prediction).

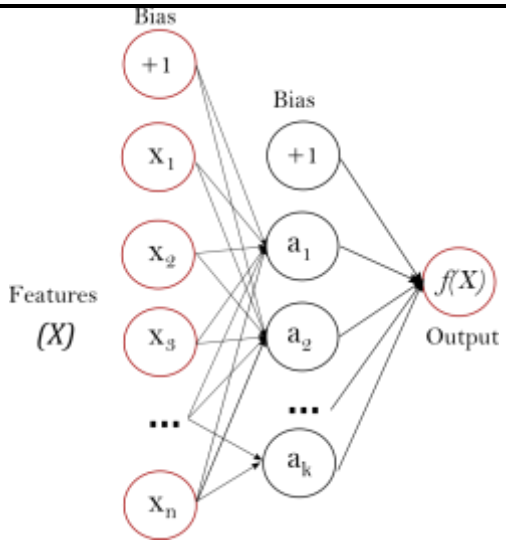

Fig. 2. *A supervised neural network learning model based on scikit learn. [16]*

Scikit-learn is primarily written in Python and heavily relies on NumPy for high-performance linear algebra and array operations. Also, to increase accuracy, some key algorithms are written in Cython. A Cython wrapper around *LIBSVM* is used to implement support vector machines; a similar wrapper around *LIBLINEAR* is used to implement logistic regression and linear support vector machines. Extending these methods with Python can be impossible in some situations. Scikit-learn works well with a variety of other Python libraries, including Matplotlib and plotly for visualisation, NumPy for array vectorization, Pandas dataframes, SciPy, and several more.

IV. SYSTEM DESIGN & ARCHITECTURE

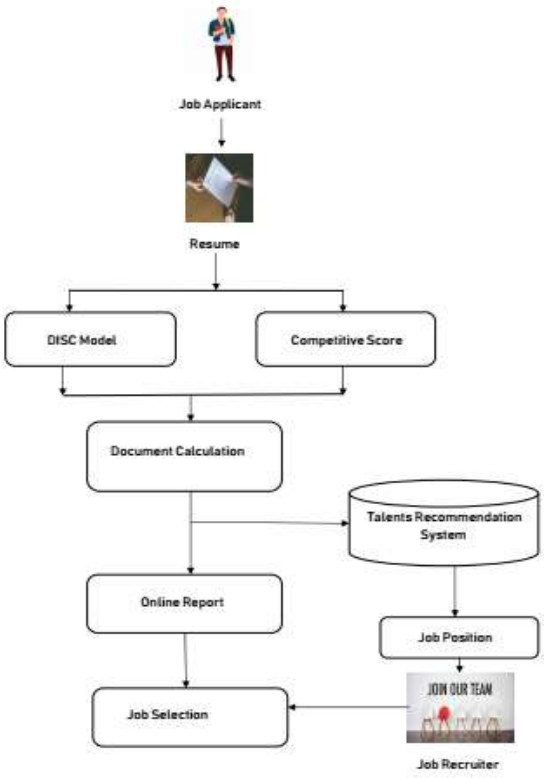

Fig. 3. *A belief overview of the system components and the architecture.*

Except for saving candidates' time by allowing them to upload their resume in whatever format they prefer, our system is able to detect all of the candidate's activity from their public network, resulting in the best candidate for that particular job, and the candidate will be satisfied because he will get a job in that company, which is reassuring. On other hand, we provide the same level of adaptability to the client business. Domain establishment, data selection, parsing, rating, and database components are among the various modules or components produced. Parsing and ranking are at the core of our framework, which was built with the Python, nltk, and tika libraries. This part performs morphological, syntactic, and semantic analysis on the candidate's data and produces parsed and ranked data based on his or her abilities. [4]

i. ***Competitive Score***

This part performs morphological, syntactic, and semantic analysis on the candidate's data and produces parsed and ranked data based on his or her abilities. This information is then stored in a database and retrieved and shown to users as needed.[5]

ii. ***Domain Establishment***

Since this proposed system is browser based and will be accessed by various users, this module is accountable for managing user accounts and databases.

iii. ***Registration/Auth Module***

If a new user decides to connect with our system, he would first signup by filling out all of the required fields, such as identification and authentication, and permissions. If the user already exists, he then must log in with valid credentials.

iv. ***Parsing and Ranking***

The Inference module is in responsibility of parsing the document and saving it in json files, which the Rankings module will then use. After that, the rankings module will leverage the json file to rate the student's results based on his or her talents, and the knowledge will be stored in a database.

v. ***Morphological Analysis***

Morphological and study of how words are constructed in natural speech in linguistics. The statement is split down into pieces, which are the smallest unit of words and establish the word's functional group.

vi. ***Syntactic Analysis***

The function of combinatorial evaluation is to understand the sentence's syntactic structure. It's also known as Hierarchical review, and it's used to detect sentences, organise tokens into grammatical phrases, and give a syntactic structure to them.

vii. ***Semantic Analysis***

The importance of creating descriptions for the meaning of language inputs is known as semantics. It discusses how to deduce the meaning of a phrase from either the value of its constituent components.

V. USER PROFILE SYSTEMS

i. Talent Recommendation System

It allows the user to change status, cancel, retrieve shortlisted candidates, update the assessment process, prepare a job description, prepare applicant specification and notify the applicant and notify the recruiter.

ii. Applicant

It performs the process of registration, searching and applying for a job, checking status and available vacancy.

iii. Recruiter

It performs viewing ranking, reports, approval, or recommendation and sends a notification to applicants.

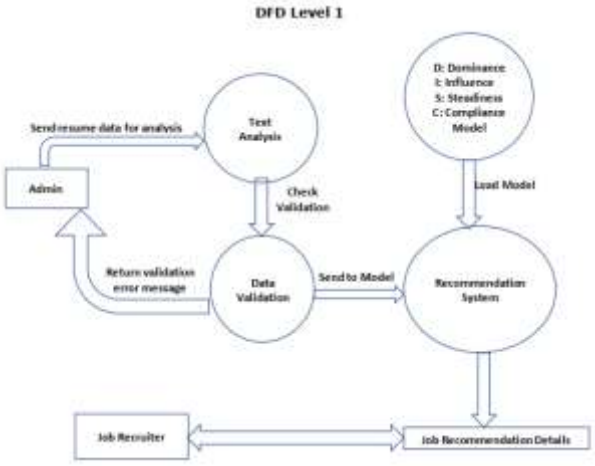

Fig. 4. *The elaborated Data Flow through the system (DFD Level 1)*

VI. PERFORMANCE MEASUREMENTS

i. Performance Evaluation Measurements

a. Cosine Similarity

Since the term frequencies are always positive, the cosine similarity equation would yield a value between 0 and 1. [15]

b. Jaccard Similarity

The aim of Jaccard similarity is to find commonality between data sets by intersecting them. The score can be calculated in Python using the Sci-Kit learn library [15]:

```
sklearn.metrics
    .jaccard_score(
        actual, prediction
     )
```

c. Perplexity

Perplexity is a numerical value assigned to each expression. To determine how reliable the NLP model is, it looks at the underlying probability distribution of the terms in the sentences. The better the formula, the lower the ranking.[15]

d. Word Error Rate

WER (word error rate) is a useful metric. It is a metric that can be used to compare two documents and is heavily dependent on the number of substitutions, deletions, and iterations between the two documents.[15]

ii. Evaluation Metrics

Task	**Description**	**Value**
Textual Analysis (NLP)	Screening the resume uploaded by the user (candidate logged in to the portal)	*< 1000ms*
Data Validation	Validating the analysed, and collected data against a semi-trained model	*500ms to 2000ms*
Calculating DiSC Metrics	Loading the DiSC model and creating a partial recommendation system based on the previous data & semi-trained models.	*< 500ms*
Calibrating & Updating Recommen-dation Systems	Analysing and modifying the recommender constantly based on collected data, various inputs and automatic model corrections.	*2 - 10 minutes per iteration (depending on the hardware)*
Recommendations for the user	Final output, Displaying users their performance and a suitable recruiter match (if any) based on their performance and skills gathered in the initial phase.	*~ 5 seconds (when data valid & model integrated)*

Table 2. *Estimated performance metrics.*

VII. ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my internal guide Prof. Ankita Karale for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful. I am also grateful to Dr. Amol Potgantwar, Head of Computer Engineering Department, Sandip Institute of Technology and Research Center for his indispensable support and suggestions.

VIII. CONCLUSIONS

Our system will provide a better and efficient solution to the current hiring process. This will provide potential candidates to the organization and the candidate will be successfully placed in an organization that appreciates his/her skills and ability. By eliminating the strenuous processes of current job applications and hiring systems, it will pave the way for a streamlined recruitment process & faster funnel processing; which in turn will reduce the expenses on hiring, onboarding, and related chain of events that currently take up unnecessary time and complexity.

IX. REFERENCES

[1] Marston, William Moulton. "Chapter VII" Emotions of Normal People. London: Routledge, 1928. 113-192. Print.

[2] Yi-Chi-Chou, Han-Yen-Yu, “Based on the application of AI technology in resume analysis and job recommendation”, 978-1-7281-3448-2/20/$31.00_c 2020 IEEE.

[3] Walid Shalaby, Bahaeddin Aila, “Help Me Find a Job: A Graph-based Approach for Job Recommendation at Scale”, 978-1-5386-2715-0/17/$31.00 c 2017 IEEE.

[4] e-Recruitment recommender systems: a systematic reviewKnowl. Inf. Syst.2021 Mauricio Noris FreireL. Castro

[5] Document-based Recommender System for Job Postings using Dense RepresentationsNAACL-HLT2018 A. ElsaftyMartin Riedl, Chris Biemann, Semantic Scholar Journal article

[6] J. Malinowski, T. Keim, O. Wendt, and T. Weitzel. Matching people and jobs: a bilateral recommendation approach. In HICSS, 2006

[7] LIONEL NGOUPEYOU TONDJI (lionel.ng.tondji@aims-senegal.org) African Institute for Mathematical Sciences (AIMS) Senegal, Supervised by: Pr. Ndeye Niang Keita Conservatoire National des Arts et Métiers, France 31 January 2018. Submitted in Partial Fulfillment of a Masters II at AIMS. "Web Recommender System for Job Seeking and Recruiting"

[8] Walid Shalaby, Bahaeddin Aila, "Help Me Find a Job: A Graph-based Approach for Job Recommendation at Scale", 978-1-5386-2715-0/17/$31.00 c 2017 IEEE.

[9] Ketki Deshpande, Shimei Pan, James Foulds, "Mitigating Demographic Bias in AI-based Resume Filtering", UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy c 2020 Association for Computing Machinery. ACM ISBN 978-1-4503-7950-2/20/07. . . $15.00 https://doi.org/10.1145/3386392.3399569.

[10] Thomas Schmitt, Philippe Caillou, "Matching Jobs and Resumes: a deep Collaborative filtering task ", HAL Id: hal-01378589 https://hal.inria.fr/hal-01378589 Submitted on 13 Oct 2016.

[11] Roshan Belsare, Dr, V. M. Deshmukh, "Employment Recommendation System using Matching, Collaborative Filtering and Content Based Recommendation", International Journal of Computer Applications Technology and Research. Volume 7–Issue 6, 215-220, 2018, ISSN:-2319-8656.

[12] Amber Nigam, Aakash Roy, Arpan Saxena, Hartaran Singh, "Job Recommendation: Leveraging Progression of Job Applications", ACM ISBN 978-1-4503-7950-2/20/07. https://doi.org/10.1145/3386392.3399569.

[13] Suhas Tangalde, Vijayaraghavan, "AUTOMATED TOOL FOR RESUME CLASSIFICATION USING SEMANTIC ANALYSIS" International Journal of Artificial Intelligence and Applications (IJAIA), Vol.10, No.1, January 2019, DOI: 10.5121/ijaia.2019.10102.

[14] Ioannis Paparrizos, Barla, Cambazoglu, Gionis, "Machine Learned Job Recommendation", RecSys'11, October 23–27, 2011, Chicago, Illinois, USA. Copyright 2011 ACM 978-1-4503-0683-6/11/10 ...$10.00.

[15] NLP: How To Evaluate The Model Performance by Farhad Malik https://medium.com/fintechexplained/nlp-how-to-evaluate-the-model-performance-7e1820cdb759

[16] scikit-learn - Machine Learning in Python https://scikit-learn.org/stable/