# Resume Analysis and Job Recommendation

Amruta Mankawade
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
amruta.mankawade@vit.edu

Vithika Pungliya
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
vithika.pungliya20@vit.edu

Roshita Bhonsle
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
roshita.bhonsle20@vit.edu

Samruddhi Pate
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
samruddhi.pate20@vit.edu

Atharva Purohit
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
atharva.purohit20@vit.edu

Ankur Raut
*Department of Artificial Intelligence
and Data Science*
*Vishwakarma Institute of Technology*
Pune, India
ankur.raut20@vit.edu

*Abstract—* **A job seeker always spends hours searching through the massive amount of recruiting information on the Internet to identify ones that are helpful. The field of job recommender systems for recruiting has just evolved and experienced rapid growth. The job recommender systems view user profiles and design recommendation technologies that have drawn interest, been studied in academia, and been put into practice for some use cases in industries. This study creates and puts into practice a recommendation system for online job searching to lessen this tedious task. In this study, an architecture has been proposed to extract the most suitable professions based on the resume of the individual. Natural Language Processing and Machine learning are used to train the model that predicts the profession. Then using this profession as a key, jobs are web scraped from Naukri.com. The required skills of the jobs are matched with the individual's skills using a cosine similarity algorithm and are ranked and displayed to the user.**

*Keywords—Resume, Cosine similarity, Naïve Bayes, Web scraping, Job Recommendation.*

## I. INTRODUCTION

With the rapid growth of the Internet came an increase in the amount of information that was accessible online, necessitating a greater need to enhance users' information management skills. This greatly piques interest in specific research fields and technological breakthroughs that could assist handle this information overload. The two most important fields are information retrieval (IR) and information filtering (IF). While IR deals with automatically matching user information IF aims to assist customers in getting rid of unwanted information.

The online recruiting platform also referred to as the e-recruitment platform, has changed how companies hire applicants and is one of the most successful business innovations [1]. Since it can be difficult for most firms to find the ideal employee, these platforms have recently grown in popularity. It has long been acknowledged that a fundamental obstacle to corporate success is the scarcity of applicants with particular skill sets. Online resources including social media applications, employment portals, and firm career web pages have been the driving forces behind this development. Employers use these portals to advertise job openings, and job seekers use them to publish their resumes.

New graduates often find themselves at crossroads while applying for new jobs. With the huge number of platforms and portals, it can be a daunting task for any new graduate to find the job they are looking for. Oftentimes such candidates are unsure of their skills and competencies and find it difficult to apply for a job that best fits their skills and experience. This is also challenging for job recruiters as they must sift through the tens of thousands of resumes received for each position that is open [2]. Although it needs human intellect to assess a candidate's suitability for a particular job profile, reviewing every application manually is not practical, and doing so might even be a waste of their time. Hence, it is becoming more and more important to use recommender system technologies to help recruiters manage this large volume of information properly.

The biggest issue is that it takes a long time to sort through all of the applications that are submitted in response to a standard Internet job posting before finding the one who is best suited for the position. A crucial stage in the hiring process is reviewing a candidate's résumé to ascertain whether or not they are qualified for a certain position. It is not practicable to manually weed out the resumes because it would take a lot of time and money, both of which the employing organizations cannot bear [3]. Additionally, this resume screening process is unfair because many qualified candidates do not receive the attention they merit. This could lead to the hiring of ineffective individuals or the passing over of qualified candidates. Consequently, by automating a component of the process, the number of candidates that a person must review can be decreased. Due to this, it has been important for a candidate to prepare his/her resume accordingly, which will differentiate it from other resumes. It will be beneficial if they already get some idea of the company they may be recruited into from their resume, and based on the results, they can work on increasing the efficiency of their cv.

With the rise of artificial intelligence and machine learning techniques, the automation of this task has been made possible. Techniques for natural language processing help interpret and parse written text to extract the necessary information [4]. Various fields such as candidate name, email address, education, and skills can be extracted from a candidate's resume using natural language processing. These skills can be extracted and compared to existing jobs to find a suitable match for the candidate. In this study, an NLP based approach is discussed to extract relevant information from the candidate's resume and suggest suitable job recommendations using Cosine Similarity and Naïve Bayes Classification. The proposed study will not only help candidates understand their own competitiveness in the industry but also recommend jobs matching the candidate's skills.

The paper is structured and organized as follows. Section II covers the literature review that contains various papers related to the existing literature and has been discussed briefly. Section III presents an in-depth discussion of the methodology of the proposed study. The results and discussions are elucidated in Section IV which summarizes the key findings of this paper. Section V is the Conclusion, and it summarizes the entire contents of the paper briefly and discusses the future scope of this study.

## II. Literature Review

[5] The authors examine the issue of employment recommendations in this research, making relevant career recommendations to the candidates based on their profiles (resume/skills). They have contrasted a baseline approach that treats users and jobs as documents and measures appropriateness with the help of cosine similarity that considers transitions in jobs and was developed using data on a group of users' career ascents. The job transition model surpasses cosine similarity. Additionally, a cascaded approach that incorporates cosine similarity and career transitions produces more recommendations of the same caliber. The analysis is carried out by looking at data from 2,400 LinkedIn members, and the approaches' accuracy in predicting individuals' present jobs based on their profiles and prior position history is assessed.

[6] The study looks at the shortcomings of current career recommendation systems, such as cold start, scalability, and sparseness. Additionally, it has been studied how suggested machine learning career recommendation system implementations bring elements of security, dependability, and transparency into the career advice process. In order to create a career recommendation system using the content-based filtering approach, it has also been investigated whether these systems can be improved.

[7] The existing hybrid recommendation systems that are model based often require a substantial amount of feature engineering to build a user profile in order to combine these two filtering processes. By directly representing the probabilistic connections between the properties of related objects, statistical relational learning (SRL) offers an easy solution for combining these two methods. Although, due to the large amount of data that is being utilized in real-world recommendation systems, there is very little research on hybrid recommendation systems using SRL models, and none of it has been applied to systems involving big data. The authors have proposed a way for developing a hybrid job recommendation system employing the SRL methods. Additionally, this approach can allow the tuning of the trade-off between not only the precision but also the recall of the system to satisfy the recommendation systems' common requirement (penalization of false positives should be harsher than false negatives since they are undesirable). The experimental findings show the effectiveness of their suggested strategy and its enhanced performance on suggestion precision.

[8] The authors of this paper proposed a model that, in essence, extracts information and statistics from a resume and ranks resumes in accordance with a company's preference and requirements using Natural Language Processing (NLP) techniques. They created a job portal where employees and candidates could upload their resumes for any available position. Using NLP technology, the required information would be parsed and a structured resume with information would be generated. Additionally, the resumes of employees would be ranked according to the skill requirements of the company and the employees' skills in the provided resume.

[9] An improved recommender system for e-recruitment is presented in this article. This system's primary function is to suggest relevant job offers that recruiters have placed to job searchers. First, web-scraping methods are used to get the data from job-related websites. To get information into a structured format, the material is cleaned and prepared using preprocessing procedures. To preserve word semantics and make similarity calculations easier, the chosen attributes are converted into vectors of real values using vectorization methods. RNN is one deep learning technology used to classify job profiles. Prior to giving an ordered list of job recommendations, scores are provided for the chosen attributes in order to improve results.

[10] This study introduces a job recommender system to match non-standard, unstructured/semi-structured job descriptions (JD) with resumes. The article first suggests a mix of NLP skills extraction strategies. On an industrial size dataset, the combined approaches' performance produced accuracy and recall values of 0.78 and 0.88, respectively. This notion of extraction of implicit skills which are not expressly listed in a Job Description (JD), but can be seen with respect to the perspective of a work role, is introduced here. Nearly, all 1.1 million JDs, obtained from web crawling were projected into this semantic space after a Doc2Vec model is trained. The abilities that are missing from a particular JD but are present in other comparable JDs are acquired, and the acquired skills are then weighted using a variety of methods to produce the final set of implicit skills. In order to compare the abilities extrapolated from the resume of a candidate to the implicit and explicit skills of JDs, a variety of similarity metrics are investigated. According to empirical data for matching resumes and JDs, the suggested methodology performs 29.4% better than a baseline technique that just considers explicit skills, with a mean reciprocal rank of 0.88.

## III. METHODOLOGY

In this section, a detailed description of the Methodology has been elucidated in the flowchart in Fig 1.

### A. Dataset

The dataset used in this proposed study is taken from Kaggle and named the "Resume dataset" [12]. It includes 962 resumes, with the resume column containing the resume in text format and the most suitable profession according to the text in the category column. It contains resumes across 25 distinct professions.

### B. Data Preprocessing

First the necessary libraries were imported. NLTK is an important library while dealing with Natural Language. A lot of the data that could be analyzed is unstructured and contains human-readable text. Before analyzing that data programmatically, it must be preprocessed. Thus, NLTK provides functions to pre-process the raw textual data. The label encoder function provided by the sklearn library is used to encode the classes in the target variable column. Spam is encoded as 1 and Ham as 0. Then the textual data is pre-processed. First, the URLs, Hashtags, Punctuations, Hexadecimal Numbers, Extra white spaces, and some common abbreviations are removed. Then the sentences are tokenized and a filtered list is created in which the stop words have been removed. By removal of these words, more focus is given to the key information in the text by eliminating the low-level information. Finally using the WordNet lemmatize, the sentences were lemmatized. The several inflected versions of a word are combined into one item for analysis by NLTK Lemmatizer. Lemmatization adds word context in a manner like stemming. As a result, it combines terms with related meanings.

### C. Model Training

The term frequency-inverse document frequency was used for converting the text into vectors that can be fed to the model. It can determine how important or relevant a document's string representations like words, sentences, lemmas, etc. are in relation to a group of documents. The IDF is needed since terms like "of," "as," "the," etc. commonly arise in an English corpus and need to be corrected. Therefore, by using inverse document frequency, the weighting of frequent terms was reduced while increasing the significance of infrequent terms. The model that was used for fitting the data was the multinomial NB model. For a given sample, the likelihood of each badge is determined, and the badge having the highest probability of all is then taken as the output. The multinomial NB algorithm was an excellent choice for the analysis of resume skills because it is primarily utilized for NLP and text data analysis [11].

### D. Testing and web scraping

A resume parser was used to extract the skills from the resume which is later fed to the model. The top 3 professions were predicted using the probabilities predicted by the model. Using these professions as the search query, jobs are scraped from a job aggregator website naukri.com. Finally, a similarity index between the skills of the user and the required skills scraped for the job was calculated using the cosine similarity algorithm. Cosine similarity is used primarily in natural language processing to compare the similarity between two text documents. To apply cosine similarity in text data, the raw text data must first be tokenized. From the tokenized text data, a similarity matrix then can be constructed and used by the cosine similarity metrics to determine how similar the text documents are to one another. The jobs were sorted on the basis of their similarity index and the top 15 jobs with the highest similarity index were displayed.
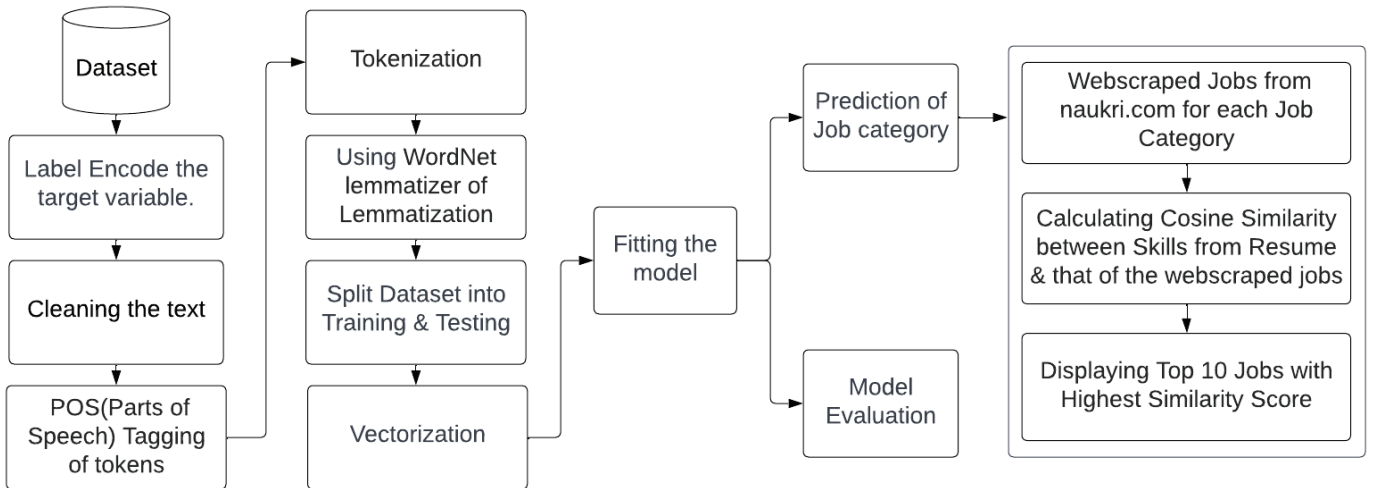


Fig. 1.  Flowchart for the Methodology

## IV. RESULTS AND DISCUSSIONS

In this section, the results of the implementation of the proposed method are discussed. Fig 2. below shows the word cloud of all resumes belonging to the Data Science class. As it is observed that words like data, science and machine learning are highlighted which means the frequency of occurrence of these words is high in Data Science related resumes.
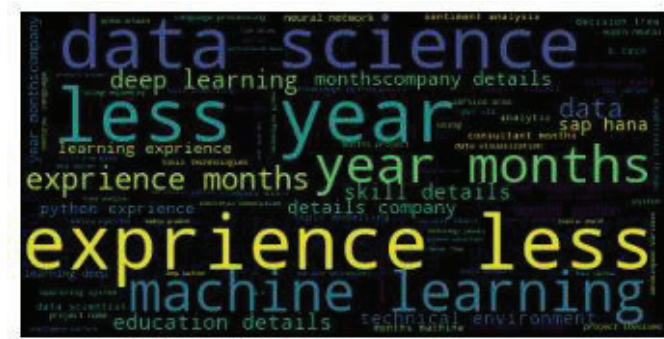


Fig. 2. Word Cloud for data science

Fig 3. Below shows the output of the trained model for a sample resume. It contains the top 3 most suitable professions on the basis of skills present in uploaded resumes.

```
The Top 3 Professions according to the Parsed Resume are:

Data Science
Python Developer
Java Developer
```

Fig. 3. Top professions list

Figure 4 Below shows the Confusion Matrix for Job Classification. All the categories are mostly predicted accurately by the proposed model.
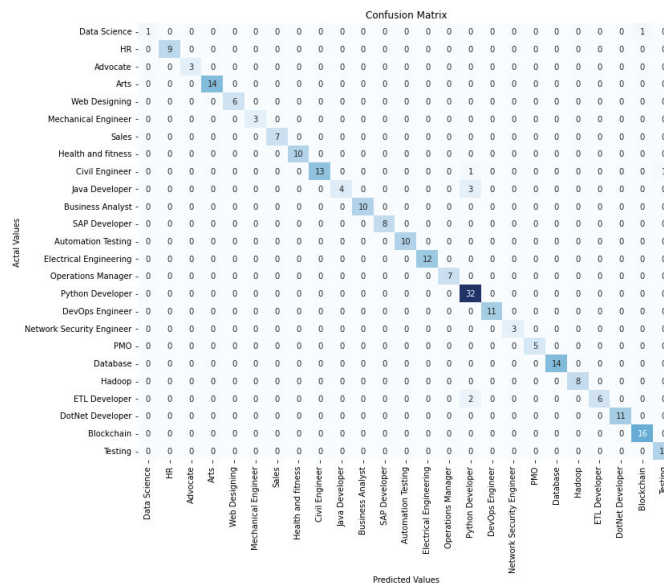


Fig. 4. Confusion Matrix for Job Classification

Table 1. Below shows the Classification Report including Precision, Recall, F1 Score and Accuracy. The model gained a Precision of 0.99, Recall of 0.95. F1 Score of 0.96 and Accuracy of 0.97.

TABLE I. CLASSIFICATION REPORT FOR RESUME DATASET

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| accuracy |  |  | 0.97 | 241 |
| macro avg | 0.99 | 0.95 | 0.96 | 241 |
| weighted avg | 0.97 | 0.97 | 0.96 | 241 |

The proposed methodology/ model was tested on another standard dataset. The dataset was for classifying emails as either spam or ham [13]. The results obtained are shown in table 2.

TABLE II. CLASSIFICATION REPORT FOR EMAIL CLASSIFICATION DATASET

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| accuracy |  |  | 0.98 | 1393 |
| macro avg | 0.97 | 0.96 | 0.96 | 1393 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1393 |

Table 3. Below shows the top 15 recommended jobs listed in a dataframe containing attributes such as Location, Rating, Salary, Job Post History, and the link to apply for the recommended job. The jobs are being sorted on the basis of cosine similarity between the skills present in the resume and the required skills for a particular job.

TABLE III. RECOMMENDED JOBS

| Sr. No. | Title | Company | Skills | Professions |
|---|---|---|---|---|
| 1. | Senior Business Analyst | Freshworks | [python],[analytical],[Business].... | Data Science |
| 2. | Analyst-Data Science | Accenture | [Consulting],[MySql],[Data Science].... | Data Science |
| 3. | Analyst, Data Science & Aanalyst | TransUnion | [SQL],[Communication],[IT Skills].... | Data Science |
| 4. | Data Science Manager | Rapido | [Data Science],[Data Analysis],[No SQL].... | Data Science |

## V. Conclusion

In this paper, the study of a system, which recommends jobs to a user according to his/her resume is presented. This system directly connects and fetches the suggested job cards from the naukri.com website, which is a platform containing all the available job profiles and is displayed to the user in the order of their relevance. For classification of resumes, Naïve Bayes algorithm has been used. The accuracy of the proposed model on testing data is 97% and the macro f1 score is 96%. For calculating the similarity index cosine similarity was used. Through this application, a job seeker can connect and search for a job suitable to his/her skills and experience. This project can be developed further in the coming time by adding the feature of candidate filtering for the companies, so that the companies can easily select suitable candidates for the job without having to go through each resume manually. This in turn will make the application useful for both job seekers and recruiters.

## References

[1] Färber, F., Weitzel, T., & Keim, T. (2003). An automated recommendation approach to selection in personnel recruitment.

[2] Chou, Y. C., & Yu, H. Y. (2020, August). Based on the application of AI technology in resume analysis and job recommendation. In 2020 IEEE International Conference on Computational Electromagnetics (ICCEM) (pp. 291-296). IEEE.

[3] Daryani, C., Chhabra, G. S., Patel, H., Chhabra, I. K., & Patel, R. (2020). An automated resume screening system using natural language processing and similarity. ETHICS AND INFORMATION TECHNOLOGY [Internet]. VOLKSON PRESS, 99-103.

[4] Sinha, A. K., Akhtar, A. K., & Kumar, A. (2021). Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review. *Machine Learning and Information Processing*, 207-214

[5] Heap, Bradford, Alfred Krzywicki, Wayne Wobcke, Mike Bain, and Paul Compton. "Combining career progression and profile matching in a job recommender system." In Pacific Rim International Conference on Artificial Intelligence, pp. 396-408. Springer, Cham, 2014.

[6] T. V. Yadalam, V. M. Gowda, V. S. Kumar, D. Girish and N. M., "Career Recommendation Systems using Content based Filtering," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 660-665, doi: 10.1109/ICCES48766.2020.9137992.

[7] Yang, Shuo, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. "Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach." *Knowledge-Based Systems* 136 (2017): 37-45.

[8] Bhor, S., Gupta, V., Nair, V., Shinde, H., & Kulkarni, M. S. (2021). Resume parser using natural language processing techniques. Int J Res Eng Sci (IJRES), 9(6), 01-06.

[9] Mhamdi, D., Azzouazi, M., El Ghoumari, M. Y., Moulouki, R., & Rachik, Z. (2020, December). Enhancing Recruitment Process Using Semantic Matching. In the International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 370-378). Springer, Cham.

[10] Gugnani, A., & Misra, H. (2020). Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(08), 13286-13293.

[11] Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., & Shashidhar, R. (2022). Movie recommendation and sentiment analysis using machine learning. Global Transitions Proceedings, 3(1), 279-284. https://doi.org/10.1016/j.gltp.2022.03.012.

[12] https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset

[13] https://www.kaggle.com/datasets/ozlerhakan/spam-or-not-spam-dataset