INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

AI-Powered Job Market Intelligence and Salary Prediction System using NLP and Machine Learning

1st Jayakumar. P
Department of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
Jayakumar456,jkp@gmail.com

2nd Dr. Vidhya Sathish
Department of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai, India

Abstract *- The exponential growth of various online recruitment sites has resulted in the creation of a large-scale heterogeneous data related to the job market, which also includes unstructured data related to the various aspects of jobs, skills, experience, and salaries offered. The analysis of the unstructured data related to the job market and the derivation of meaningful insights from the same is a challenge for the conventional data analysis techniques. This paper proposed an AI-based "Job Market Intelligence and Salary Prediction System" using various "Natural Language Processing" and "Machine Learning" techniques for the effective analysis of the unstructured data related to the job market. Different text preprocessing and TF-IDF techniques have been effectively implemented by the proposed system for the effective transformation of the unstructured data related to the job market. A supervised learning-based model has also been implemented for the proposed system for the salary prediction using various input features for the proposed system.*

Introduction

The increased popularity of online Recruitment websites has led to the generation of to jobs, such as job descriptions, skills required for the jobs, level of experience, and salaries for the jobs [1][4]. Processing such heterogeneous data using traditional methods is challenging due to their unstructured nature and complexity [10]. Recently, Artificial Intelligence (AI), specifically Natural Language Processing (NLP) and Machine Learning (ML), has been proposed as a solution for obtaining useful information from such data [3][11].

The proposed project aims to develop a new AI-based Job Market Intelligence System for Salary Prediction using NLP methods such as Text Pre-processing and TF-IDF Vectorization for transforming the data into a more structured numerical format [8][9]. The proposed system also uses a Supervised Learning approach with Random Forest Regression for Salary Prediction based on job roles, skills, and experience [5]. Moreover, the proposed system also provides a facility for Resume-Job Matching using Cosine Similarity and Skill Gap Analysis [6][17]. The proposed system aims to help job seekers take better decisions for their career growth and enhance their employability skills [19][20].

A. Contribution of this paper

- The paper proposes a new AI- based Job Market Intelligence and Salary Prediction System using NLP and Machine Learning techniques for efficient analysis of job market information.
- The paper proposes a new approach to convert unstructured job market information and resumes into efficient numerical information using text pre-processing and TF-IDF techniques.
- The paper proposes a Random Forest Regression algorithm to accurately predict salaries using job roles, skills, and experiences.
- The paper also proposes useful features such as resume-job matching, skill gap analysis, and job trends using this proposed system.

Literature Survey

Job portals and recruitment sites are becoming increasingly popular. These days, many organizations post their job details on these. The volume of data is seeing an increase. The employment market now requires computational intelligence methods. A machine learning approach to the job market trends is presented in the paper.

The style of data analysis is shifting from numeric to textual. Why so? Analyzing text using regular software is not effective. The text analysis is required for job market analysis. Job descriptions, Resumes, Skills…etc are referred to as text.

Breiman [5] developed a random forest regression model for job classification and salary prediction. To meet the task, researchers proposed other models as machine-learning evolved. Still, random forest turned out to be relatively better for it. It can cope with complex relations, and prevents us from overfitting. Due to the same logic, salary prediction prefer ensemble models over any independent

we made a resume-job matching system.

model. For example, linear regression and SVM [10]. The classification of job-posting data into their respective job fields is a very simple yet important task which is being used by many systems and organizations. This is the basis of the paper. Different organizations solicit huge job postings, which need to be processed and added to the existing database. Many recruitment systems and platforms have employed various methods MORE recently.

Cosine similarity can also be used to match a resume with a job description and other similar tasks. For example, using cosine similarity, a recommender system can recommend suitable jobs to a candidate in accordance with his resume [16] [17]. In the same vein, a skills gap analysis determines which skills the candidate lacks when applying for a specific job. Recent publications also shed light on the utilization of extensive datasets of job descriptions, resumes, or job titles sourced from LinkedIn, Kaggle, and similar platforms aimed at analyzing fascinating trends in employment [19][20]. We can witness the fact.

Proposed Methodology

The suggested system will gather data regarding jobs online from job portals or datasets. This data usually consists of job descriptions, skills, experience, salaries, etc. Since our data is unstructured, we will apply various NLP processing methods such as tokenization, stop-word removal, and text cleaning.

Once the preprocessing is over, we use the TF-IDF vectorizer to create a feature set in numbers. Eventually, the prepared data remains available for training a supervised learning model. Our system is built using Random forest regression through which salary will be predicted based on job role, skills, years of experience etc.

Cosine similarity is used for matching the candidate to a specific job. Cosine similarity measures the similarity between a candidate's resume and a job description. Our enterprise has implemented a skill gap analysis module which helps to identify the missing skills by comparing the profiles of the candidates.

The system uses a system on the interface using Streamlit. So, any user can easily access the system to input data and get the prediction and information.

Architecture Diagram

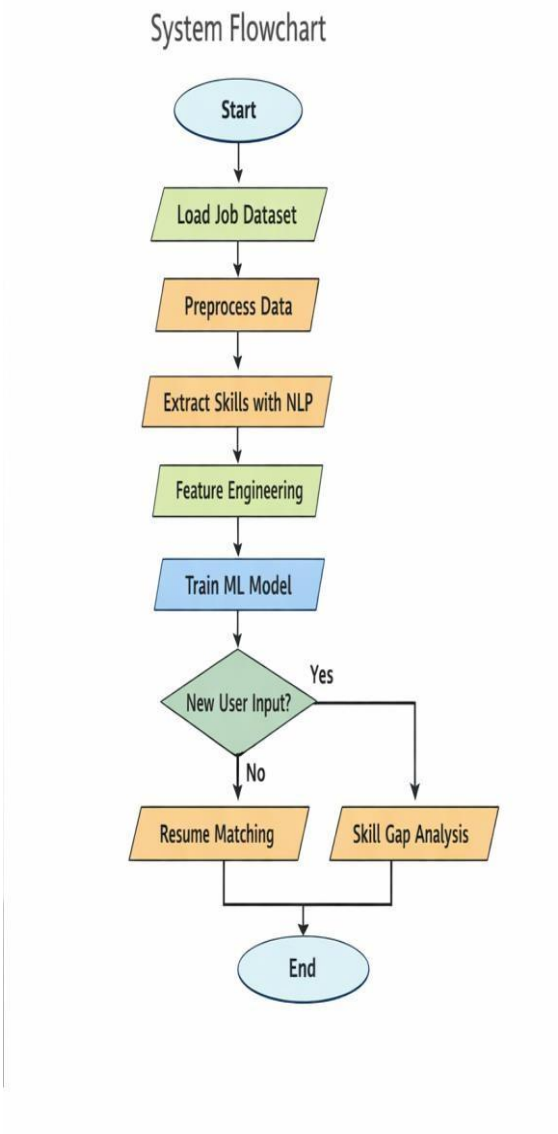


Fig.1 data flow diagram

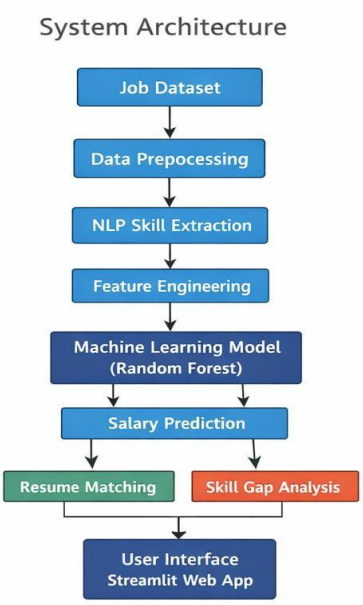


Fig.2 Architecture diagram

Various Phases / Methodologies

- Data Collection

Job feature extraction is the process of collecting information about job listings found on job boards or in datasets. Common examples of job features that can be collected include job titles, job postings, key words associated with a job, minimum and maximum years of experience required, as well as salary ranges.

- Data Preprocessing

All the data that is being collected is unstructured. So it needs to be processed. Most of the time this processing includes removing unwanted characters, converting to lowercase, tokenization and removing stop words.

required. Typically sourced from job boards or labour market datasets.

- Candidate Resume Data

- Feature Extraction

Data cleaning has to be done for all the data that is to be analysed. The data has to be converted to a numerical format by using a technique called TF-IDF in order to understand the importance of the words or the features.

- Model Building

In this phase, we built a Machine learning model using Random Forest Regression to train a model that can predict the salaries based on the job title, required skills and years of experience.

- Resume-Job Matching

Resume matching with cosine similarity against job descriptions.

- Skill Gap Analysis

This step finds the gaps in the skill set of a candidate for the job. It is very helpful to understand what needs to be improved in a candidate's skill set in relation to a job advertised.

- Job Market Analysis

This system analyses employment trends for individual skills and occupations, including required skills and average salary.

- System Implementation

All the above functionalities are combined into a web application using Streamlit.

- Evaluation and Testing

We now have all parts in place and can finally think about evaluating the performance of our system. The ultimate goal is to measure how accurate the predictions of salary values for job openings and corresponding candidate profiles are.

Input

- Job Description Data

Job description data is a sub-set of job market data which provides information on a job role such as: job role description, job responsibilities, key skills and qualifications - User resumes/profiles including skills, education, experience & projects.

- Skills Information

Skills are extracted from the job ads and the resumes. The skills extraction process allows for the comparison and analysis of the skills from the two datasets.

- Experience Details

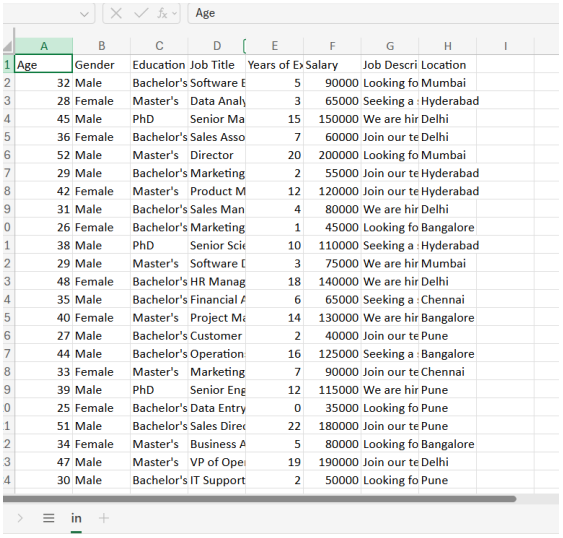
The years of experience required for this job and the years of experience provided by the candidate.

- Job Role / Title

Job title for the position, e.g. Software Engineer or Data Analyst.

- Salary Data (Training Input)

Information about previous salaries that have been included in datasets can be used to train a model developed using machine learning.



	A	B	C	D	E	F	G	H	I
1	Age	Gender	Education	Job Title	Years of Ex	Salary	Job Descri	Location	
2	32	Male	Bachelor's	Software E	5	90000	Looking fo	Mumbai	
3	28	Female	Master's	Data Analy	3	65000	Seeking a	Hyderabad	
4	45	Male	PhD	Senior Ma	15	150000	We are hir	Delhi	
5	36	Female	Bachelor's	Sales Asso	7	60000	Join our te	Delhi	
6	52	Male	Master's	Director	20	200000	Looking fo	Mumbai	
7	29	Male	Bachelor's	Marketing	2	55000	Join our te	Hyderabad	
8	42	Female	Master's	Product M	12	120000	Join our te	Hyderabad	
9	31	Male	Bachelor's	Sales Man	4	80000	We are hir	Delhi	
0	26	Female	Bachelor's	Marketing	1	45000	Looking fo	Bangalore	
1	38	Male	PhD	Senior Scie	10	110000	Seeking a	Hyderabad	
2	29	Male	Master's	Software D	3	75000	We are hir	Mumbai	
3	48	Female	Bachelor's	HR Manag	18	140000	We are hir	Delhi	
4	35	Male	Bachelor's	Financial A	6	65000	Seeking a	Chennai	
5	40	Female	Master's	Project Ma	14	130000	We are hir	Bangalore	
6	27	Male	Bachelor's	Customer	2	40000	Join our te	Pune	
7	44	Male	Bachelor's	Operation	16	125000	Seeking a	Bangalore	
8	33	Female	Master's	Marketing	7	90000	Join our te	Chennai	
9	39	Male	PhD	Senior Eng	12	115000	We are hir	Pune	
0	25	Female	Bachelor's	Data Entry	0	35000	Looking fo	Pune	
1	51	Male	Bachelor's	Sales Direc	22	180000	Join our te	Pune	
2	34	Female	Master's	Business A	5	80000	Looking fo	Bangalore	
3	47	Male	Master's	VP of Oper	19	190000	Join our te	Delhi	
4	30	Male	Bachelor's	IT Support	2	50000	Looking fo	Pune	

Implementation

The proposed system is built on Python with some machine learning and NLP Libraries to extract meaningful description. Another major aspect that is covered by the system is the skill gap analysis where the system identifies the job description to highlight which sections and keywords in a resume match to the job.

information about the job market. Firstly, the job description, required skills, years of experience and salary are loaded from the created dataset. A data preprocessing step is then required to ensure the quality of the data. Preprocessing is carried out on the loaded data to convert all the text to lowercase, remove special characters and stop words.

Now we will apply the TF-IDF vectorization on the data after preprocessing. Machine learning algorithms cannot directly classify raw text data. So we have to convert text data into some numerical data. We will then split our dataset into training set and test set.

This application applies a Random Forest Regression model to the data provided and hence uses the training data to train the model with job descriptions, key skills and years of experience. With this trained model the application will then predict a salary given any new input.

Cosine similarity is used to match resumes to jobs. It will show how the resume and job description match up for the candidate and which job they will be the best fit for. A skill gap analysis is also done by matching the candidate's skills to the job requirements and highlighting gaps.

Now, we will combine all the skills we learned so far to create a fun tool to calculate potential salaries based on given information and then see how well the resume matches the job postings based on the skills they ask for. We'll use Streamlit to create the user interface for the web application. Have a look at the final product and then we'll break it down.

Output

Output Features The system is capable of outputting multiple values based on the input given by the user. This will include the estimated salary of a particular job based on the required skills and the candidate experience as well as a matching score between a resume and required skills which needs to be learned or improved by the candidate. This is finally given as an output in an easily interpretable manner.

Result and Discussion

So as to validate the whole system, we carry out the evaluation on a database of jobs that correspond to diverse occupations, required skills and salaries. In this case, the results obtained demonstrate that the salaries predicted by the Random Forest Regression algorithm are acceptable especially when the data is large and accurate. In addition, the resume-job matching part using the cosine similarity to obtain the most relevant matches for the candidate resumes with respect to the job descriptions also works as expected.
Skill gap analysis showed us which skills the candidates need to work on. Apart from that, the system is doing exactly what it should – using NLP and machine learning to derive useful information from unstructured job descriptions. While the system isn't always 100% accurate, it's usually due to the poor quality or insufficient amount of the job data.

Screenshots / Charts / Graph

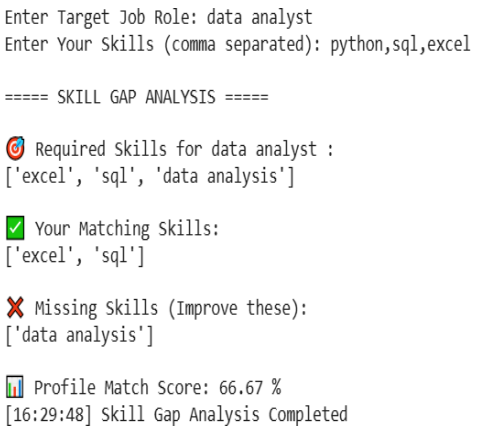


Fig.3 Skill gap analysis

job description to highlight which sections and keywords in a resume match to the job.

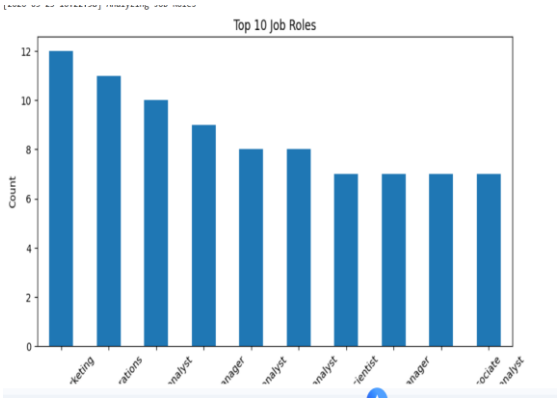


Fig.4 Bar graph

Fig.7 Salary prediction

Fig .5 Correlation matrix

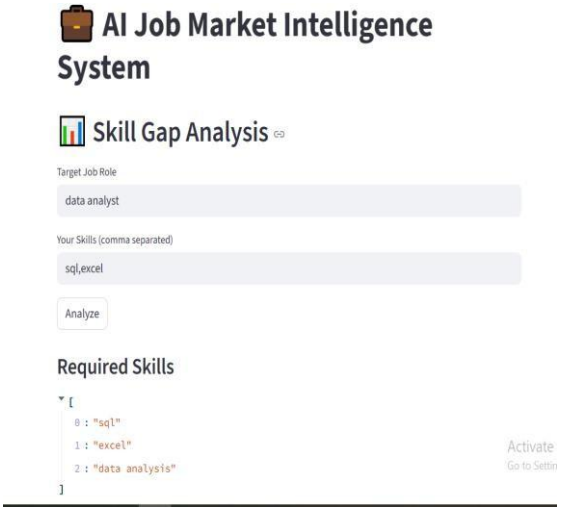


Fig.8 Streamlit webpage

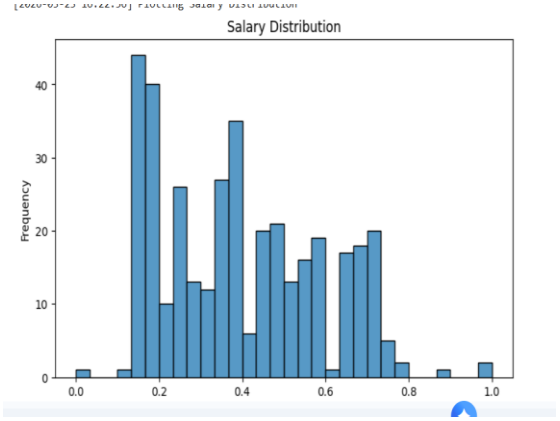


Fig.6 Stacked bar graph

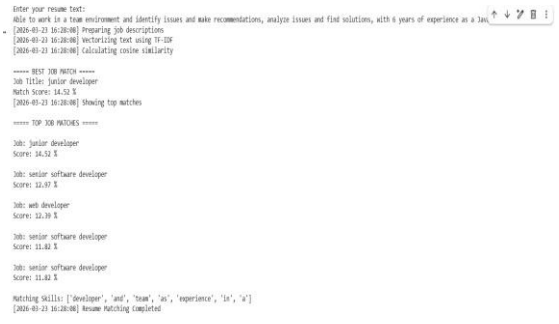


Fig.9 Resume match

Conclusion

The “Job Market Intelligence and Salary Prediction using AI” project has reached the completion milestone. The resulting AI

application will be contributing significantly in supporting job seekers in their job search and their career development. The application is built on Natural Language Processing (NLP) and machine learning technology, and is equipped with functions such as salary prediction, resume matching and skill gap analysis.

The results obtained are very promising and we foresee a high number of applications in the labour market to understand the needs of the market regarding the required skills and salary for the different job positions. Our research can be a basis or reference for the workers, in order to make them aware of their role and the necessary improvements and training needed in terms of required skills and salary, in order for the worker to be aware of the demanded expectations. Our system may help to improve the employability of the workers and to have a better data driven career planning. On the next iterations, we will proceed to deepen in the following points: More information More models to do the deep learning in the process Extraction of skills.

Refrences

[1] Mikolov, T., et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781.

[2] Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. EMNLP.

[3] Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[5] Breiman, L. (2001). *Random Forests*. Machine Learning Journal, 45(1), 5–32.

[6] Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR, 12, 2825–2830.

[7] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

[8] Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*. Information Processing & Management.

[9] Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*.

[10] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

[11] Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer.

[12] Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press.

[13] Liu, B. (2020). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

[15] Zhang, Y., & Wallace, B. (2015). *A Sensitivity Analysis of CNNs for Sentence Classification*. EMNLP.

[16] Resnick, P., & Varian, H. R. (1997). *Recommender Systems*. Communications of the ACM.

[17] Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. Springer.

[18] Carvalho, V. R., & Cohen, W. W. (2005). *On the collective classification of email "speech acts"*. SIGIR.

[19] Kaggle (2023). *Job Salary Prediction Dataset*. https://www.kaggle.com

[20] LinkedIn Economic Graph (2022). *Global Talent Trends Report*. LinkedIn Report