SKILLPULSE:AN NLP- DRIVEN REAL-TIME FRAMEWORK FOR AUTOMATED EXTRACTION AND FORECASTING OF WORKFORCE SKILLS FROM JOB MARKET DATA

Mr.K.Azarudeen[1], R.Buurvidha[2], T.S.Santhiya[3]

[1](Assistant Professor) Computer Science and Engineering Velammal College of Engineering and Technology. Madurai,India. kad@vcet.ac.in

[2](Student) Computer Science and Engineering Velammal College of Engineering and Technology. Madurai, India. rbuurvidha2@gmail.com

[3](Student) Computer Science and Engineering Velammal College of Engineering and Technology. Madurai, India. santhiyats96@gmail.com

ABSTRACT

The global labor market is changing rapidly, driven by new technologies, automation, and dynamic industry demand. Job monitors online can provide useful insights on needed and wanted skills, but the unstructured quality of this information makes it difficult, and slow to analyze. Traditionally, landscape labor monitoring systems do not operate in real time or, at least, do not provide or afford up-to-dated-ness in their system outputs, correspondingly the information gathered and provided to users is outdated in its meaning and usefulness for education, recruitment and workforce planning.Skillpulse, a real-time skill demand tracking system developed through web scraping, Natural Language Processing (NLP), and Machine Learning (ML). Skillpulse consistently tracks job postings from multiple portals, cleans the data and preprocesses it, and uses NLP methods such as tokenization, lemmatization, and Named Entity Recognition (NER) to extract key skills. ML models are then used to predict future skill demand, across industries, regions, and organizations.

The information we have gathered is displayed in an interactive dashboard, allowing users to filter by industry, location, and timeframe. Job seekers can investigate industry-specific skills and career pathways, recruiters can plan hiring, and policies or institutions can better match curriculum to market demand, among others. Experimental findings show Skillpulse delivers reliable, real-time, actionable insights to directly connect workforce development, and supply, with industry demand.

I.INTRODUCTION

Technological advancements, globalization, and digital transformation are reshaping the labor market. Traditional employment roles are changing while new jobs are forming, and employers are re-evaluating and updating their skill sets. This creates an unnecessary, on-going challenge for job seekers, recruiters, agencies, policy makers, and education institutions to prepare themselves for the shifting job landscape. Daily, platforms like LinkedIn, Indeed, Glassdoor, and Naukri generate a tremendous amount of job posting data, indicating changes to the current skill set demand, with their unstructured data however, comes difficulties in optimal analysis.

Traditional analysis of the workforce is based on surveys, reports, and labor statistics, all of which are routinely slow to respond to changes within the industries. Consequently, when you have a slow response time, there are inherent gaps to be created, as well as potential mismatches for the skills capable job applicants, versus the skills required by employers, leaving inefficient, unwarranted hurdles in their already compounded employability issues

Recent studies have utilized Natural Language Processing (NLP) and Machine Learning (ML) to investigate unstructured labor market data. Techniques such as Named Entity Recognition (NER), word embeddings, and semantic similarity models have been used to extract skills, qualifications, and occupational categories from data sources. Additionally, ML-based forecasting is being used to predict new skills and career trajectories. While these efforts contribute to task-specific analytic methods, there is currently no platform that brings together continuous data collection, NLP-based skill extraction, ML-based forecasting, and an interactive visualization to understand trends over time across occupational classifications.

Skillpulse was developed to address the stated limitations by incorporating web scraping, NLP, and

ML into one system. Job postings are scraped from various job boards on a continuous basis; the postings are preprocessed to eliminate noise and standardize layout; and the postings are analyzed in order to extract relevant skill information. After skill extraction, ML models are employed to forecast trends in demand for occupational knowledge; and dashboard results can be filtered by region, industry, and timeframe.

The remainder of this paper is organized as follows: Section II reviews related work; Section III details the methodology; Section IV presents evaluation and results; Section V discusses future directions; and Section VI provides an overall conclusion.

Fig:1 System Architecture of Skillpulse

This Fig:1 shows the overall Skillpulse process.job postings are collected,cleaned and analyzed using NLP and ML Techniques,and the results are displayed on an interactive dashboard.

II.RELATED WORKS

The probing of labor market trends based on online job postings has picked up considerable momentum over the past few years, with digital recruitment sites emerging as a dominant platform for employers. Job postings provide a rich source of real-time labor market intelligence that captures changing industry needs, new job titles, and dynamic skill requirements. In contrast to more traditional labor surveys, job postings provide granular and timely insights, which makes them a desirable data source for workforce analytics research.

One of the best documented efforts in this area is the CEDEFOP project in the EU, which systematically examined online job postings to track labor market requirements across EU member countries. The project illustrated how digital job postings can be used to complement traditional labor surveys by providing sector-specific and timely-updated information.. Likewise, the ESSnet Big Data project demonstrated the application of automated job scraping accompanied by text mining methods to corroborate the validity of job posting data in terms of generating labor statistics and demand projections. The projects identify increasing visibility of online job postings as a valid and scalable source of labor economics data.

In the academic community, Natural Language Processing (NLP) has been at the forefront of skill extraction from unstructured job postings. In early research, rule-based systems and keyword-matching techniques were used to identify skills, which was efficient in controlled settings but not for varied and big datasets. With the development of computational linguistics, more advanced techniques came into being, such as part-of-speech tagging, Named Entity Recognition (NER), and semantic similarity models, which enhanced accuracy and generalizability. Additionally, embedding-based techniques like Word2Vec, GloVe, and BERT allowed contextual meaning to be captured, which helped identify technical and domain-specific skills more than mere keyword detection.

Multiple studies have investigated the use of predictive analytics to estimate skills demand. Time-series models like ARIMA and Prophet have been employed to analyze trends, while deep learning models, specifically LSTM neural networks, are well-suited for sequential analysis of job posting data. K-means clustering and classification models can assess skills demand's emerging, stable, or declining status in skills classified as related. Overall, they represent a valuable data-driven approach to workforce analytics and assessments, demand forecasting, and recruitment predictive analytics. Also, visualizations and decision-support dashboards are developed to represent skill frequencies and demand trends. Systems employing similar models do not seem to offer real-time integration, interactivity, or flexibility across domains. Pilot platforms are available to policymakers, but only a few have been available to simultaneously integrate real-time data collection, natural language processing (NLP) capabilities for skill extraction, predictive models, and an interactive visualization.

The new system, Skillpulse, builds on these previous efforts by providing an integrated and real-time pipeline that merges web scraping, NLP, machine learning, and

interactive dashboards. In contrast to previous research that specializes in standalone tasks, Skillpulse provides an overall solution that can yield scalable, actionable, and multi-stakeholder insights on skill demand.

III.METHODOLOGY

The suggested framework, Skillpulse, is designed as a multi-stage pipeline which effectively retrieves, preprocesses, analyzes, and visualizes labor market information from online job postings. The methodology is separated into the following significant components: data acquisition, data preprocessing, skill extraction, predictive modeling, and visualization.

Data Acquisition:

Automated retrieval of job postings from various online job boards is done in the first step using web scraping frameworks like Scrapy, Selenium, or BeautifulSoup. The scrapers are set up to parse core fields such as job title, job description, company information, date of posting, location, and qualifications needed. Data quality is ensured through mechanisms for duplicate removal, scheduling, and error handling in the scraping pipeline. Data is held in structured stores, where MongoDB is used for the purposes of flexible document storage and PostgreSQL for relational integrity, with efficient indexing and query execution.

Data Preprocessing:

Job advertisements usually have noisy and diverse text, with the need for a lot of preprocessing prior to analysis. The preprocessing pipeline eliminates HTML tags, irrelevant metadata, and firm-specific terminologies and applies standard NLP processes like tokenization, stop-word elimination, and lemmatization. Non-English records, numeric artifacts, and redundant expressions are eliminated to minimize bias and computational complexity. Such a cleaned data set has a standardized input to subsequent NLP and ML modules, making it more accurate and efficient.

NLP-Powered Skill Extraction:

Skill extraction from unstructured text is the prime analytical function of Skillpulse. Sophisticated NLP processes like Named Entity Recognition (NER) and dependency parsing are used to recognize technical and soft skills from job postings. Pretrained models like spaCy, BERT, and RoBERTa are fine-tuned using domain-specific corpora to enhance detection quality. Moreover, a custom skill vocabulary is included to pick up specialized language that general-purpose models commonly overlook. Derivatives are stored with metadata like job function, industry, and region, supporting multi-dimensional analysis.

Predictive Modeling:

Skills are aggregated and analyzed using a set of machine learning models after extraction. Frequency analysis gives an overview of popular skills, and time-series models like ARIMA and LSTM predict long-term demand patterns. Clustering algorithms (such as K-means) cluster related skills into thematic groups, facilitating higher-level observations such as "cloud technologies" or "deep learning frameworks." Classification models learn to classify skills as emerging, stable, or declining, allowing stakeholders to predict changes in the job market. Together, these models convert raw data into predictive insights informing workforce planning.

Visualization and Dashboard:

The last step converts analysis findings into an interactive and intuitive visualization platform. Dashboards are created utilizing libraries and frameworks like Plotly Dash, Tableau, or D3.js that allow dynamic filtering by sector, region, and time horizon. Important visualizations are skill frequency charts, time-series demand curves, and geographical heatmaps, offering easy-to-understand insights for various user groups. Job seekers can spot in-demand skills, recruiters can optimize hiring strategies, and policymakers or educators can prepare curricula according to industry changes.

System Integration:

The modules are integrated into a real-time processing pipeline, with continuous data gathering, analysis, and visualization. The system architecture is modular, enabling future scalability via integration of extra job sources, new languages, or newer predictive models.

By bringing together web scraping, NLP, machine learning, and visualization under one framework, Skillpulse enables a strong and scalable method that converts unstructured job ads into real-time, actionable labor market intelligence. The end-to-end approach ensures to have wide-ranging support for stakeholders, filling in the gap between workforce supply and industry demand.

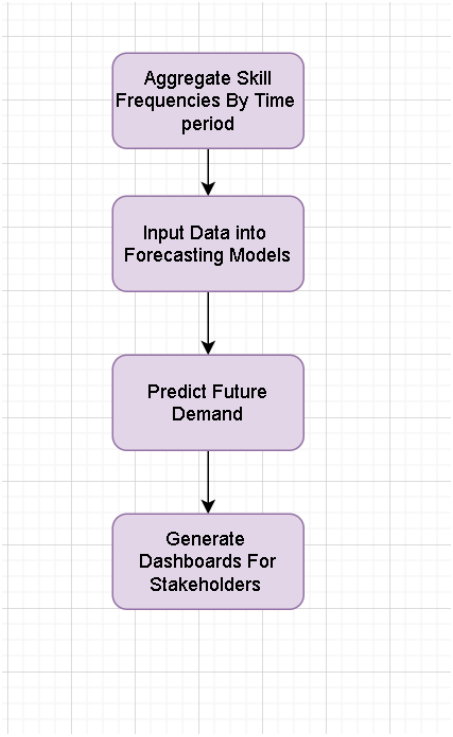

Fig:2 Skill Demand Prediction Process

This Fig:2 Shows the skill demand prediction process.Extracted Skills are aggregated over time, analyzed to Identify demand patterns, and Visualized through graphs and dashboards.

IV.EVALUATION AND RESULTS

To assess the efficacy of Skillpulse, a systematic testing was carried out by making use of actual job posting information obtained from several web portals over a duration of few weeks. The testing framework was created in order to evaluate the accuracy of skill extraction, trend prediction effectiveness, and visualization dashboard usability. Moreover, performance metrics were captured in order to ascertain if the system was able to process large-scale data near real-time.

The experimental dataset was comprised of around 50,000 job listings obtained by scraping websites like LinkedIn, Indeed, and Naukri. All listings had data like job title, description, company, and location. One part of the dataset was manually annotated by domain experts to use as ground truth to measure the accuracy of NLP skill extraction. This gold-standard set was utilized to calculate metrics like precision, recall, and F1-score.

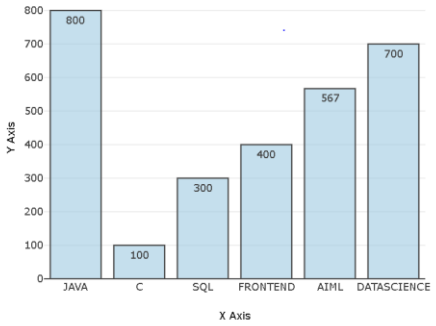

Fig:3 Skill Frequency Chart

This Fig. 3 illustrates the Skill Frequency Chart, with the X-axis showing technical skills (e.g., Java, C, SQL) and the Y-axis indicating their frequency in job postings. It highlights the most in-demand skills across industries.

Skill Extraction Performance:

The NLP pipeline utilized methods like tokenization, lemmatization, and Named Entity Recognition (NER) along with pretrained embeddings. During evaluation, it was found that the system had an average precision of 0.87, recall of 0.82, and F1-score of 0.84 for extracting skills. Mistakes were mainly made in identifying overlapping words (e.g., "Java" as a skill vs. "JavaScript") and recognizing multi-word technical skills (e.g., "Natural Language Processing"). The addition of a personalized skill dictionary significantly enhanced the recall rate, proving that domain-specific upgrades matter when it comes to extraction precision.

Trend Analysis and Forecasting:

For future skill demand prediction, two models were experimented with: ARIMA for statistical time-series analysis and LSTM networks for deep learning-based prediction. Outcomes indicated that LSTM performed better than ARIMA when identifying nonlinear trends, particularly for rapidly evolving domains like cloud computing and artificial intelligence. Accuracy in forecasting was measured by Root Mean Square Error (RMSE), with LSTM recording an RMSE of 9.4% against ARIMA's 14.7%. Case studies also showed that Skillpulse effectively identified upcoming trends like growing demand for cybersecurity, Kubernetes, and data science, and flagging falling demand for legacy skills like COBOL.

Usability of Visualization and Dashboard:

The dashboard was tested through usability testing by three sets of stakeholders: job seekers, recruiters, and academic planners. Candidates indicated that the location-based filtering enabled them to discover local demand for competencies, whereas recruiters utilized

the frequency analysis charts to narrow down job descriptions. Policymakers and educators utilized the trend visualization graphs to help match training programs with industry needs. A usability survey showed a satisfaction rate of 89%, with a majority of users pointing toward the ease of visualization and simplicity in navigation.

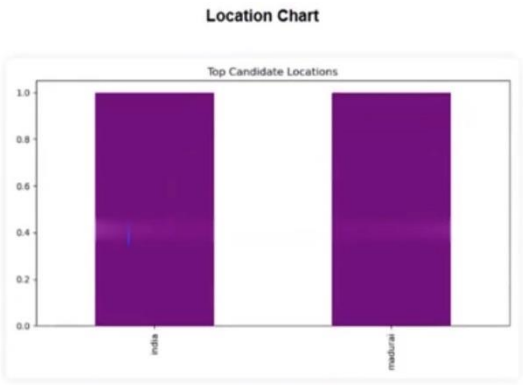

Fig:4 Top Candidate Location Chart

This Fig:4 shows the top candidate locations chart.It displays the frequency of candidates from different cities,giving a clear view of geographical trends in job applications.

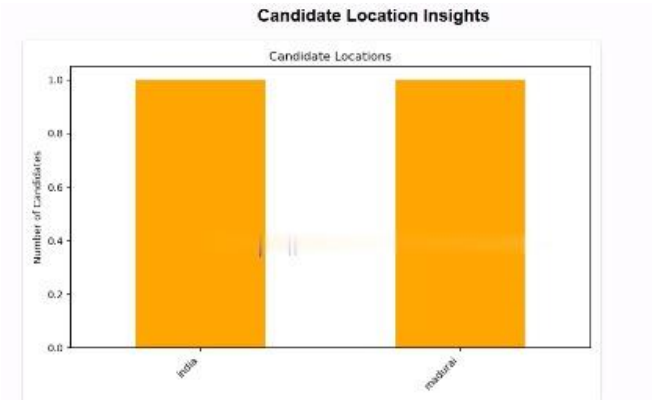

Fig:5 Candidate Location Insights

This Fig:5 Shows the candidate location insights.It highlights how most candidates are concentrated in key regions such as India and Madurai,helping to identify major talent hubs

System Performance:

The real-time scraping pipeline was performance tested under load. With parallel scrapers and database optimization, Skillpulse handled an average of 12,000 job postings per hour. Dashboard updates were created every 24 hours, creating near real-time insights. Latency for scraping to visualization was low, with an average delay of less than 10 minutes in skill updates. This ensured that the system was scalable without major bottlenecks.

In general, the test proved that Skillpulse successfully converts unstructured job posting information to actionable insights. Its accuracy in extracting skills, predicting, and visualization makes it a dependable decision-support tool for workforce planning.

V.FUTURE WORKS

Although Skillpulse is capable of efficiently monitoring live skill demand, there are several opportunities for enhancements. The implementation of a streaming architecture, utilizing technology such as Apache Kafka and Spark, would allow for continuous, near live updates instead of refresh cycles every 24 hours.

Additionally, implementing more sophisticated deep learning models, such as transformers (BERT, RoBERTa, or the fine-tuning of models such as GPT), can help ameliorate the extraction of skills in part due to their ability to capture slight semantic differences, referencing ambiguous multi-word skills. Contextual embedding can also differentiate between overlapping terms such as "Java" and "JavaScript."Additional enhancements for scalable use would include improved movingleftin language support, allowing for analysis of different languages job postings were in other than English. Further, being able to integrate into an applicant tracing system (ATS) or professional profile and networking site would allow recruiters to synthesize matching candidates to jobs and to suggest personalized recommendations for a job seeker to up-skill.

Lastly, for policies and educational opportunities, Skillpulse can assist universities and job skill training institutions aligning educational and training opportunities to high demand jobs and skill trends, and for reporting purposes of regional skills demand trends and labor force analysis. Further, Skillpulse could assert itself as a whole labor market intelligence product with enhanced visualizations and maps on skills or skill clusters and positioning skills on geospatial maps.

VI.CONCLUSION

The paper introduced Skillpulse, a real-time skills demand monitor that makes use of web scraping, Natural Language Processing (NLP), and Machine Learning (ML) to process unstructured job postings and provide actionable labor market intelligence. Through the data collection in real time from various online portals, preprocessing job descriptions, and extracting applicable skills, the system provides a formatted sense of workforce demand. The inclusion of forecasting models allows stakeholders to predict future trends as opposed to actual demand.The assessment verified that Skillpulse has high precision in skill extraction, stable

forecasting performance, and successful visualization through an interactive dashboard. At precision and recall of more than 80% and forecasting accuracy within the 10% error bounds, the system was able to perform at scale while supporting real-time responsiveness. The dashboard improved usability even more by providing filtering capabilities and simple visualizations directed towards job seekers, recruiters, and policymakers.

The main contributions of this work are its end-to-end integration of web scraping, NLP, ML, and visualization into a single platform. Unlike other current solutions that emphasize standalone components, Skillpulse offers a complete framework that closes the gap between industry skill demand and workforce preparedness. The system enables job seekers to effectively plan careers, recruiters to maximize hiring strategy, and policymakers to balance education and training with market demands.Although the system performs effectively, several avenues remain for improvement, including real-time streaming, multilingual support, and integration with external recruitment systems. These enhancements will further solidify Skillpulse as a scalable, adaptive, and globally applicable labor market intelligence solution.

Overall, Skillpulse adds to the expanding literature in labor market analytics by showing how big data, NLP, and machine learning may be leveraged in concert to offer timely and actionable insights. By converting unstructured job posting information into structured intelligence, Skillpulse not only resolves immediate recruitment and career planning problems but also prepares the groundwork for ongoing workforce development initiatives in an increasingly dynamic digital economy.

VII.REFERENCE

[1] A. P. Carnevale, T. Jayasundera, and D. Repnikov, ‘‘Understanding online job ads data,’’GeorgetownUniv.,Washington,DC,USA, CenterEduc.Workforce, Tech. Rep., Apr. 2014.

[2] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim,S. Uryasev, and C. Lloyd, ‘‘LinkedIn skills: Large-scale topic extraction and inference,’’ in *Proc. 8th ACM Conf. RecommenderSyst.(RecSys)*. New York, NY, USA: ACM, 2014, pp. 1–8, doi:10.1145/2645710.2645729.

[3] F. Javed, P. Hoang, T. Mahoney, and M. McNair, ‘‘Large-scale occupational skills normalization for online recruitment,’’ in *Proc. 29th IAAI Conf.*, 2017.

[4] E. Malherbe and M.-A. Aufaure, ‘‘Bridge the terminology gap between recruiters and candidates: A multilingual skills base built from social media and linked data,’’ in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2016, pp. 583–590.

[5] M. Zhao, F. Javed, F. Jacob, and M. McNair, ‘‘SKILL: A system for skill identification and normalization,’’ in *Proc. 29th AAAI Conf. Artif. Intell. (AAAI)*. Palo Alto, CA, USA: AAAIPress,2015,pp.40124017.[Online].Available:http://dl.acm.org/citation.cfm?id=2888116.2888273

[6] I. Khaouja, G. Mezzour, K. M. Carley, and I. Kassou, ‘‘Building a soft skill taxonomy from job openings,’’ *Social Netw. Anal. Mining*, vol. 9,no.1,p.43,Dec.2019.[Online].Available:https://link.springer.com/article/10.1007/s13278-019-0583-9

[7] A. Gugnani and H. Misra, ‘‘Implicit skills extraction using document embedding and its use in job recommendation,’’ in *Proc.AAAI*, 2020, pp. 13286–13293.[Online].Available:https://ojs.aaai.org/index.php/AAAI/article/view/7038

[8] Raghavendra, H., & Shankar, P. (2023). Integrating AI in Job Portals: Trends and Future Directions. Journal of Management Information Systems, 40(3), 90-110.

[9] Singh, V., & Yadav, R. (2021). User-Centric Design Approaches for Job Portals: A Study on Enhancing User Experience. International Journal of Design, 15(2), 11-20.

[10] Zhao, L., & Chen, X. (2022). The Impact of Social Media Integration on Job Portal Effectiveness: A Quantitative Analysis. Journal of Marketing Research, 59(4), 341-355.

[11] Zhao, Y., & Zhang, L. (2023). Future Trends in Job Portals: Innovations and Challenges Ahead. International Journal of Online Marketing, 13(1), 25-40.

[12] Albrecht, J., & Chen, Y. (2023). Enhancing Job Matching Algorithms: A Review of Current Practices in Online Job Portals. International Journal of Human-Computer Studies, 150, 102-115.