International Conference on Business Research
University of Moratuwa, Sri Lanka
December 3, 2021, 143- 152

BUSINESS RESEARCH UNIT
FACULTY OF BUSINESS
UNIVERSITY OF MORATUWA

INFORMATION EXTRACTION FROM SRI LANKAN JOB ADVERTISEMENTS VIA RULE-BASED APPROACH

R.M.H.D. Bandara[1], H.A.S.S. Gunasekara[2], W.A.D.S. Peiris[3], W.M.H.C. Wijekoon[4], T.S. De Silva[5], S.G.S. Hewawalpita[6] and H.M.S.C. Rathnayake[7]

Emails: [1]harini.17@business.mrt.ac.lk, [2]suwani.17@business.mrt.ac.lk, [3]diluni.17@business.mrt.ac.lk, [4]himali.17@business.mrt.ac.lk, [5]tilokad@uom.lk, [6]supungs@uom.lk, [7]samadhic@uom.lk

ABSTRACT

One of the major problems in the Sri Lankan labour market is the lack of availability of demand side information. This lack of information has created a gap between supply and demand of labour. Job advertisements provide a wide range of real-time information about aspects, such as skills and qualifications, that are in demand, though this information is largely unstructured and exists in many different formats. The objective of this research is to create a structured dataset of job vacancies in Sri Lanka using publicly available job advertisements. A total of 3500 images of job advertisements were scraped from Sri Lankan English newspapers and job websites and converted into text form using Optical Character Recognition (OCR). Next, a structured dataset was created by extracting information, applying a rule-based approach in the Natural Language Processing (NLP) domain, after which some basic insights on the labour market were derived. The creation of this kind of dataset could provide huge value to employers, job seekers and policymakers, providing up-to-date information on the skills and qualifications required in the job market.

Key Words: NLP, OCR, Information Extraction, Job advertisements, Labour market intelligence

1. Introduction

Skill mismatch is considered a major problem in the labor market, resulting in underemployment or unemployment. The skill mismatch occurs when there is a disparity between skills demanded and skills in supply. This mainly results due to the lack of knowledge and information available for both companies and job candidates.

The job market is a highly volatile market; the jobs in demand, qualifications and skills needed change rapidly. In order to stay ahead, both employees and employers need to pay attention to these changes in the market. While traditional studies of labor demand have used surveys of employers, data on vacancies and job advertisements are increasingly being used to measure labor demand, given that many job postings are now hosted online (Kureková et al.,2015). Moreover, problems that are common with surveys, such as inconsistencies in the data, inaccurate or inadequate responses and respondents providing misleading information due to social desirability, are avoided through this approach. Indeed, Pedraza et al,(2019), who compared vacancy information collected by using a survey method with information obtained by scraping the internet, found that the scraped information was more detailed and up to date(Pedraza et al., 2019).The use of AI algorithms to study the labour market by classifying online job vacancies(Boselli et al., 2017) resumes(Li et al., 2017) has now become a prominent method.

In the Sri Lankan context, the mismatch between labour demand and supply is a topic of high relevance, with youth unemployment in the 2020 first quarter is 26.8% and the highest rates of youth unemployment among the most educated groups (DCS 2020). At present, most national level labor market analysis is conducted by the Central bank and the Department of Census and Statistics using questionnaires and surveys, though information on labour demand is collected very infrequently (Sri Lanka labour demand survey 2017, Employability Skills Development Programme 2018). In addition to that most of the studies were conducted based on the secondary data that are almost published reports. In addition to that most of the studies were conducted based on the secondary data that are almost published reports. International Labor Organization (ILO) surveys, Consumer Finances and Socio-economic Surveys (CFSES) conducted by the Central Bank of Sri Lanka (CBSL), Population Census (PC) and Labor Force Surveys (LFS), Land and Labor Utilization Survey, Labor Force and Socio-economic Surveys conducted by the Department of Census and Statistics of Sri Lanka (DCSSL) are some of the published and data surveys that most of the research papers has addressed (Don Karunaratne, n.d.)To the best of our knowledge, most research on labour demand in Sri Lanka has not exploited job advertisements as a source of data given that the data is largely unstructured and available in a variety of formats.

This research aims to develop a structural dataset that can be used to derive insights on job market demand in Sri Lanka through information extraction from two data sources: web portals and newspapers. Given that most job advertisements were published as images, the first step, objective 1 of the analysis involved scraping 3500 images of job advertisements from English newspapers and job websites and converting them into text form using Optical Character Recognition (OCR). Next, objective 2, a structured dataset was created by extracting information, applying a rule-based approach in the Natural Language Processing (NLP) domain. A number of insights on labour market demand were derived from this dataset.

Having this kind of data source with real time information that candidates and employees alike can refer to would be beneficial for bridging the labour market mismatch. Job candidates would be able to optimize their skill- set, learn new skills required by employers and maximize their value. This would also be beneficial for policymakers and academic institutions to ensure that the labour supply meets these job market trends.

2. Literature Review

With increasing cognizance of the internet, the number of job vacancies advertised in websites are expanding. However, the history of using job advertisements as a data source for deriving insights on the labor market goes back to the 1970s (Harper, 2012). The major motivation of job posting studies are to identify changes in natural skills required by the job market (Colombo et al.,2019, Nasir et al., 2020).

Natural Language Processing (NLP) enables us to extract and identify the patterns in the textual data. Automated information extraction through text-based data has been implemented in various sectors such as project management, construction industries and so on (Zhang & El- Gohary, 2016). As many job advertisements are posted online (close to 70% of job openings in the country, according to Jayasundera et al. (2014)), real-time data has become readily available. Thus, labor analytics companies, recruitment agencies and job search engines are carrying out many researches using these advertisements as a data source (Javed et al., 2015).

A job advertisement mainly consists of job title, requirements and responsibilities. Hence, the two main tasks in creating a structured dataset are to extract the job postings and then the required skills. In Kurekovv et al. (2012), an analysis of job advertisements was done to identify skills and personal attributes that are in demand. Litecky et al., (2010) used text and web mining techniques and created a dataset of 244,460 job advertisements and titles in the US where ad requirements were extracted by pre-defining keywords. Prabhakar et al. (2005), conducted research to identify changing demand for skills considering online job advertisements in the US, examining whether the advertisements contained any of the pre- specified 59 keywords. Margareth et al (2017) created an ontology-guided demand analysis using a self-defined, Skills and Recruitment Ontology (SARO), where named-entity tagging was used by linking skills using co-word analysis.

Most of these studies used predefined keywords or dictionaries such as International Standard Classification of Occupations (ISCO) (Pedraza et al. 2019, Sebastiani, 2002). Wowczko (2015) used a different approach, where descriptions were analyzed, and words were reduced until only significant ones remained. There are also a few researchers who have used a manually designed class set - Loth et al. (2010) used a text-mining approach of bag-of- words, mapping all text to a classifying space that is common, while Kessler et al. (2010) created separate Character n-grams or groups of words.

In conclusion, with the rapid increase in data and technology, a new research era has elevated the analysis of labor markets. In considering the related literature, many countries have attempted to analyze the demand side of the labor market through web based data, specially the job adverts, most of these researches analyze job adverts for a specific industry (Brancatelli et al. 2020, Lovaglio et al. 2018). Given that labour market demand research in Sri Lanka is still conducted in the traditional way, this research will

use job advertisements related to the entire job market to analyse real-time labour market demand. Considering all the points mentioned above, this research can be recognized as novel and insightful.

3. Methodology

Figure 1 shows the process that is being followed to achieve the two objectives of this research.

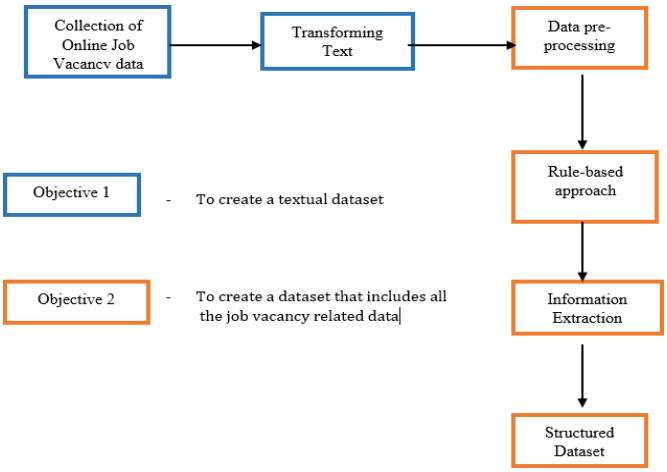

Figure 1: Methodology

Source: Author constructed

Data Gathering:

At present, many online sources are available for job searching like e-papers and online job portals. The first objective was creating a textual data set using online job advertisements. Since using only one type of information source leads to biases (Kureková et al. 2015) multiple sources were used, namely online newspapers (e- papers) and online job portals. Sunday Observer and its sister newspaper Daily News were selected as newspapers due to the free availability of data on the respective websites, while Top Jobs was chosen as the online job portal, given that it includes advanced labor market information system data and has partnerships with around 1500 leading organizations in Sri Lanka.

The python library Beautiful Soup was used to scrape the job advertisements from Top Jobs. However, the advertisements in Sunday Observer and Daily News were scraped manually as the websites are designed in a manner where automatic scraping is difficult. In total, 3535 job advertisements were scraped using these sources. Then the image data set was transformed into textual format. Google API was used in order to create the textual dataset.

Information Extraction:

The aim in the information extraction phase is to extract desirable information through unstructured data and convert it to a structured data format. The created structured data can be used to derive insights. There are two main approaches to information extraction: the rule- based approach and the machine learning approach (Tang et al., 2007). Rule based approach can be identified as one of the oldest and widely used methods of extraction and work best on predefined patterns in the data that need to be extracted. The machine learning approach involves working with developing machine learning algorithms to understand predefined textual data sets and training to extract the desired information. One of the advantages of the rule-based method is ease of implementation compared to the machine learning approach though the variety of information captured in this method is limited as this approach only works for explained rules.

In this study, we have only used the rule based approach to extract information. The fields that were to be extracted for this study were identified as job title, company name, educational qualifications, years of experience, salary, age limit, location, address, gender preferences (if any), telephone number and email.

The data pertaining to the identified fields were extracted using regular expression and phrases- matching techniques. Regular expression methods can be identified as pattern identification methods. In this methodology a regular pattern of how the data can be apprised is fed into the computer. For example, telephone numbers have ten digits and start with zero (0) or (+94). Phrase matching techniques are used to extract the set of words from a given word when a specific rule is not available, but the appearance of an interesting phrase follows the same pattern in the data set. For example, “LTD” and “PLC” are always preceded by the company name. Therefore, the company name was extracted through this approach. For some attributes if we are certain of the outcomes the dictionary phrase matcher is used. For example, gender is an attribute we can predefine the outcome as male, female, lady and gentlemen. This approach was tested on gender, education qualification and locations and job title extraction. The accuracy of the information extraction can be improved through using larger dictionaries and a wide array of customized patterns.

4. Results

Table 1 shows the coverage percentage and the accuracy of using the rule based approach. The coverage percentage was calculated by dividing the number of extracted results by the number of advertisements. Accuracy of the produced rules were assessed using the following method. 100 advertisements were chosen randomly and the accuracy of the extracted information was assessed manually. If the extracted information matched the real information, a score of 1 was given, if not, a score of 0 was given. The accuracy percentage figures were calculated by dividing the number of accurately extracted results by the number of extracted results. There were several limitations in extracting these entities as some patterns were not available in our pattern database, some advertisements were not properly converted and some advertisements were unique and represented information in a unique way. This limitation was especially prevalent for extraction of job title and company name. A limitation in extracting the location was that most extraction libraries such as geotext only extract cities and not sub cities. Since there are only a few major cities in Sri Lanka, only those were extracted using the libraries.

Table 1: Coverage and Accuracy percentages of the entities

Entity	Coverage Percentage	Accuracy Percentage
Job title	68%	56%
Company Name	65%	56%
Education Level	84%	82%
Years Experienced	37%	59.45%
Salary	62%	90%
Age Limit	70%	70%
Location	73%	80%
Address	74.60%	61.70%
Gender	77%	96%
Telephone Number	82%	100%
Email	93.33%	100%

Source: Author constructed

Preliminary labor market insights:

The following categories are in accordance with the categories specified in top job web portal.

Table 2: Top 5 job categories in the dataset

Category	No. of advertisements
IT-Sware/DB/QA/Web/Graphics/GIS	582
Sales/Marketing/Merchandising	315
Eng-Mech/Auto/Elec	144
Office Admin/Secretary/Receptionist	144
Accounting/Auditing/Finance	138

Once the structured database on job advertisements was created, labor market insights were derived using descriptive statistics and graphical methods. Majority of advertisements were for IT related jobs. A total count of 1808 company names was extracted and KPMG Executive Search (Private) Ltd was the firm which posted the highest number of job advertisements. Out of 3510 job ads extracted 2109 required some qualifications educational, professional or both.

Figure 2: Word Cloud for qualification

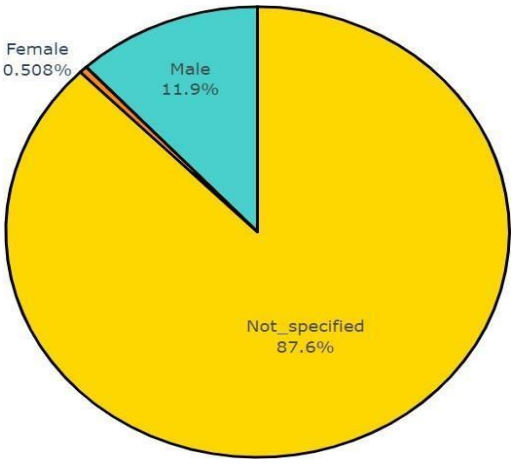

Figure 3: Gender division

Most demanded educational qualification is a degree and CIMA (Chartered Institute of Management Accounting) is the most demanded professional qualification. The IT-Sware/DB/QA/Web/Graphics/GIS category posted the majority of job advertisements and the most demanded education qualification was a degree and the most demanded professional qualification was CIMA, even though Accounting/Auditing/Finance was the 5th most posted category and the reason for this is that the jobs in the first four categories that have posted the majority of advertisements do not require professional qualifications (refer Table 2).

A gender disparity can be seen in the job market (Figure 3). This is a huge social issue. Jobs advertisements in search of jobs such as teachers and receptionists specifically ask for female candidates while jobs requiring night shifts, delivering goods, and needing technical hardware knowledge demand male employees. One main reason for such disparity is the social norm. There is a social view that some specific jobs are more suitable for females and some are more suitable for males. Having gender-coded job roles will subconsciously affect the recruitment process significantly.

The average age of an employee requested by employers according to the advertisements is 37 years old and the average years of experience needed is 3 years. For further analysis, a dashboard was created using Power Bi. Readers can access the dashboard using the link below. Insights On Labor Market.pbix

5. Conclusion and Implication

This study uses web scraping, OCR and NLP to create a structured database of job vacancies in Sri Lanka using online job advertisements. Using rule-based methods for information extraction allowed us to achieve coverage of over 71% and accuracy of over 77% for our selected fields. This structured database was then used to derive some insights about labor demand in the country. Furthermore, the average age limit was 37 years and on average 3 years of experience was needed. It is expected that this type of analysis can be useful for job- seekers as well as employers, educational institutions and policymakers to reduce the labor market mismatch in the country.

This analysis has some limitations. First, the analysis is based on English Job advertisements, which may create a bias towards a certain type of high-skilled jobs. We

had to limit ourselves to English job adverts because no proper OCR tool is created yet to identify Sinhala or Tamil text from images, however we recommend that if such a tool can be developed a wider range of jobs can be extracted which could contribute to derive better insights. Another limitation in this study is that the sources are limited to one web portal and one newspaper. Only Sunday Observer e-paper was freely available online other newspapers were not freely available to scrape data. This research can be further expanded using data from other web portals in the country and social media sites such as LinkedIn. Therefore, further studies can be carried out using more data sources. Finally, we only use rule-based methods for information extraction in this paper, which is less suited for data where patterns are not predefined; this limits the type and accuracy of information that can be extracted. In future work, we plan to manually annotate the data and apply machine learning models to further improve the accuracy of the information extracted and, thereby, the quality of the structured dataset.

References

Brancatelli, C., Marguerie, A., & Brodmann, S. (2020). Job creation and demand for skills in Kosovo: What can we learn from job Portal Data? https://doi.org/10.1596/1813-9450-9266

Colombo, E., Mercorio, F., & Mezzanzanica, M. (2019). AI meets labor market: Exploring the link between automation and skills. Information Economics And Policy, 47, 27-37. doi: 10.1016/j.infoecopol.2019.05.003

Don Karunaratne, H. (n.d.). *Structural Change and the State of the Labour Market in Sri Lanka**

F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao and T. S. Kang, "Carotene: A Job Title Classification System for the Online Recruitment Domain," 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015

Harper, R. (2012). The collection and analysis of job advertisements: A review of research methodology. Library And Information Research, 36(112), 29-54. doi: 10.29173/lirg499

Jayasundera, T., Repnikov, D., & Carnevale, A. (2014). The Online College Labor Market: Where the Jobs Are. Georgetown University Center.

Kureková, L., Beblavý, M., & Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal Of Labor Economics*, *4*(1). doi: 10.1186/s40172-015-0034-4

Kurekovv, L., Beblavy, M., & Haita, C. (2012). Qualifications or Soft Skills? Studying Job Advertisements for Demand for Low-Skilled Staff in Slovakia. SSRN Electronic Journal. doi: 10.2139/ssrn.2402729

Kessler, R., Torres-Moreno, J.-M., & El-Beze, M. (2010). E-gen : Traitement automatique d'informations de ressources humaines. *Document Numérique*, *13*(3), 95–119. https://doi.org/10.3166/dn.13.3.95-119

Loth, Romain & Battistelli, Delphine & Chaumartin, Francois-Regis & de Mazancourt, Hugues & Minel, Jean-Luc & Vinckx, Axelle. (2010). Linguistic information extraction for job ads (SIRE project). RIAO 2010.

Lovaglio, P. G., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(2), 78–91. https://doi.org/10.1002/sam.11372

Litecky, C., Aken, A., Ahmad, A., & Nelson, H. (2010). Mining for Computing Jobs. IEEE Software, 27(1), 78-85. doi: 10.1109/ms.2009.150

Margareth, E., Scerri, S., Morales, C., Auer, S., & Collarana, D. (2017). Ontology-guided Job Market Demand Analysis: A Cross-Sectional Study for the Data Science field. Proceedings Of The 13Th International Conference On Semantic Systems. doi: 10.1145/3132218.3132228

Nasir, S., Wan Yaacob, W., & Wan Aziz, W. (2020). Analysing Online Vacancy and Skills Demand using Text Mining. Journal Of Physics: Conference Series, 1496, 012011. doi: 10.1088/1742- 6596/1496/1/012011

Pedraza, P., Visintin, S., Tijdens, K., & Kismihók, G. (2019). Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data. *IZA Journal Of Labor Economics*, *8*(1). doi: 10.2478/izajole-2019-0004

Prabhakar, B., Litecky, C., & Arnett, K. (2005). IT skills in a tough job market. Communications Of The ACM, 48(10), 91-94. doi: 10.1145/1089107.1089110

Pedraza, P., Visintin, S., Tijdens, K., & Kismihók, G. (2019). Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data. IZA Journal Of Labor Economics, 8(1). doi: 10.2478/izajole-2019-0004

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

Sri Lanka labour demand survey 2017. Department of Census and Statistics, Ministry of National Policies and Economic Affairs, 2017.

(2018). Retrieved 25 August 2021, from https://slasscom.lk/wp-content/uploads/2019/10/Survey-on-employablility-skills-2018.pdf

Tang, Jie & Hong, Mingcai & Zhang, Duo & Liang, Bangyong & Li, Juanzi. (2007). Information extraction: Methodologies and applications. Emerging Technologies of Text Mining: Techniques and Applications. 10.4018/978-1-59904-373-9.ch001.

Wowczko, I. (2015). Skills and vacancy analysis with data mining techniques. Informatics, 2(4), 31–49. https://doi.org/10.3390/informatics2040031

Zhang, J., & El-Gohary, N. (2016). Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. Journal Of Computing In Civil Engineering, 30(2), 04015014. doi: 10.1061/(asce)cp.1943-5487.0000346