Analysing Online Vacancy and Skills Demand using Text Mining

S A Md Nasir¹, W F Wan Yaacob², and W A H Wan Aziz³

¹,²,³Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
Kelantan, Bukit Ilmu, Machang, Kelantan, Malaysia

¹syerina@uitm.edu.my

Abstract. The issue of graduates' employability has become a major concern in many countries as the number of graduates keep increasing over the years. Fresh graduates are experiencing difficulties in finding meaningful employment as the growth in hiring has slowed in the recent period. With the impact of fourth industrial revolution, today graduates face a world transformed by technology. The role of universities in higher education become more challenging in shaping more digital technology graduates with the right skills and knowledge. Problems can be seen at both management level of higher education in producing graduates who can meet the needs of industry while industry is also facing the problem of finding skilled graduates who suit their needs. To examine the work skills demanded from Industry of Malaysian, this study proposed a text mining approach from 465 unstructured data of analyst job vacancies in classifying job titles and extracting vacancy and skills from job descriptions. The methodology involved are text retrieval, text pre-processing, word cloud corpus generation and text clustering. The findings indicate that the selected job posts emphasize more on vacancies information rather than skills needed which further shows the gap existence between the skills demand and supply.

1. Introduction

The problem of graduates’ employability remains a national continuing issue and putting pressure for universities to improve graduate employment outcomes. According to Malaysian Ministry of Education’s Graduate Tracer Study, of the 273,373 graduates in 2015, a large number hold Bachelor’s Degree and Diploma, 45% and 43% of all graduates, respectively. Among all graduates, 53% were reported to have started working, 18% chose to pursue further studies, and 24% of graduates were still unemployed. The high level of youth unemployment is caused by too many people graduate from university with a qualification that does not get them a job. Thus, higher education plays an important role in providing a trained workforce. Among several key policy questions being raised are the types of jobs being created and the readiness of the human capital base, and measures to enhance matching in the labour market and alleviate information asymmetry on industry skill needs. However, graduates are not aware of specific competences of job descriptions from the industries that evidence about the skills needed. Minimalizing mismatch between skills needs and supply is now become a concern of policy makers around the world.

Most of the studies investigate the attributes that contributes to successful employment of graduates [1-3] and identifies important competencies for improving graduate employability [4]. Though more parameters have been mentioned, there is still an issue of mismatch between the real data of qualification and skills needed by the industry that is not addressed thoroughly. Currently, the online labour market
is a powerful force and offers great insight into selected aspects of employment. A practical approach to analysing skills demand from online vacancies is made possible through text mining which able to retrieve and analyse the content of job advertisements. In this paper, vacancy and skills analysis will be performed on online job posts in order to determine the vacancy and skill sets needed by the industry particularly in this case, analyst job vacancies. Therefore, the objectives of this study are 1) to scrap specific competences from job descriptions and 2) to segment the job descriptions through text clustering. A brief review of the literature is presented on approaches and techniques of data mining and text mining. Next, the steps taken for text mining approach are discussed. Then follow with the results of the analysis and finally discussed the implications for the future of research on reducing the gap between skill demands and supply.

2. Related works

The persistence of the graduate unemployment problem warrants the need for predicting factors that are associated with low employability graduates from higher institutions. This issue can be solved by developing a predictive model to predict potential graduates’ employability. There were a wide range of prediction models that has been developed by previous researcher including Probit and Tobit models [5] during early twenties. The models developed only used historical data on graduates’ employability traditional parameter such as socio-economic conditions, academic performance, ethnicity and type of degree. However, recently as the world transformed by technology with availability of large amount of data, the development of the prediction models in predicting graduates’ employability is moving from the traditional statistical model to data mining approach via machine learning [1, 6-8]. More parameters were used to predict graduates’ employability that includes additional skill such as cognitive, higher order thinking, technical, personal, practical, generic and others.

Jantawan & Tsai [8] develop a model to predict employees’ performance by applying data mining techniques specifically Bayesian method and the Tree method. Work done by [9] use actual data from the graduates themselves after six months of graduation in order to identify attributes that influenced graduates’ employability in profiling them as employed, unemployed or others using various algorithms under Bayes and decision methods. A study by Daud et al. [10] predict whether a student will be able to complete his degree or not through learning analytics, discriminative and generative classification models. Attributes such as family expenditures and student’s personal information also influenced on student performance prediction. On the other hand, Thakar & Mehta [6] proposed a unified predictive model by integrating clustering and classification techniques to provides a generalized solution for student employability prediction.

Apart from that, analysing massive online data has gained popularity among researchers recently through text mining techniques. Study done by [11] examined Business Intelligence (BI) and Big Data (BD) competencies by performing an automated content analysis of job ads through text mining approach. The findings reveal that the demand for BI competencies is still far bigger as they found three times more job ads than for BD competencies. According to [12], by using text mining approach, the study able to identify and characterize the skills and competencies currently required in the workplace which is needed to inventory their existing workforce skills in order to identify critical future human resources. Meanwhile, [13] applied text mining approach specifically Latent Dirichlet Allocation (DLA) topic model to identify skill sets from online job advertisements. [14] reported that Latent Semantic Indexing (LSI) model proposed in this study produces much finer and more timely readings of the labour market compare to the common official employment or job opening statistics which can be applied to different job markets, commercial sectors and geographical regions.

Employers these days are progressively worried about finding appropriate workers with specialized skills as well as furnished with high level state of employability skills and capacity to modify with fast changes in the business [15-16]. Online recruitment offers many benefits since many websites are being created, updated, and actively promoted and hence, produces massive amounts of vacancy data and provides volumes of potentially useful information that can be collected and analysed [17]. Digital
information has grown exponentially which resulted in sharp rise in demand for Data Scientist, Data Analyst, Data Artists and Data Visualiser vacancies, all over the world [18]. A report from the recent Digital Workforce of The Future by LinkedIn, which revealed an increase of 21% growth in demand in Malaysia comes from combination of skills encompassing Big Data, data analytics and web development [19]. Hence, this study will focus on a single segment which is analyst vacancies. It was also acknowledged that Analyst jobs are the most diversified, highly demanded and dynamic group, therefore it was particularly interesting for further exploration.

3. Methodology
This study adopted text mining approach by examining the content of job advertisements on analyst vacancy based on two types of information: job titles and job descriptions. Data published from Jobstore.com domain between May and June 2019 was retrieved using Import.io online app, a sample of 465 Analyst vacancies. An open source machine learning and data visualization, Orange software is used for data analysis purposes. The methodology comprises four phases namely; Text retrieval, text pre-processing, word cloud corpus generation and text clustering.

3.1. Text retrieval
The first stage in text retrieval involved obtaining the textual data. The data used for this study was retrieved from Jobstore Malaysia website, (https://www.jobstore.com/my/search-jobs). A sample post is shown in figure 1.

![Figure 1. Sample of the document retrieve from the website.](image)

This website contains vacancy and skill for analyst offered in Malaysia. This study used selected document regarding job title and description for analyst job posted from April 2019 until June 2019. The two months’ time framework is chosen to capture the latest posting on job vacancies for analyst. A total of 465 documents were extracted from 1,720 analyst job vacancies.

3.2. Text Pre-processing
In Orange software, a widget called preprocess text is used to clean the dataset of analyst which obtained from online website (see figure 2). Preprocess text splits the text into smaller units known as tokens. Steps involved in the analysis are as follows:
4

Figure 2. Preprocess Text parameters.

- **Information on processed data**: 465 number of documents on the input, 19653 total number of tokens in the corpus and 2363 number of token types in the corpus.
- **Transformation**: turn all text in the corpus to lowercase.
- **Tokenization**: Regexp with \w+: matches only words is selected.
- **Filtering**: Stopwords removes stopwords from text (e.g. removes ‘and’, ‘or’, ‘in’...). In this study, Extra Stopword.txt file, own list of stopwords with one stopword per line is loaded to help remove unnecessary word.

3.3. **Word Cloud Corpus Generation**

Word cloud corpus can be generated with an add-ons option in Orange software. Word Cloud widget is connected after Preprocess Text widget in order to find the most frequent words in the text. Word Cloud displays tokens in the corpus together with the frequency of the word in corpus denoted as word size. A list of words presented by their frequency (weight) in the widget. The widget outputs documents, containing selected tokens from the word cloud which will be assigned with random colour.

3.4. **Text Clustering**

To perform text clustering, there are few steps needed to be carried out after cleaning the data using **Preprocess Text** widget (refer Figure 3 for further details). The **Bag of Words** model creates a corpus with word counts either absolute, binary (contains or does not contain) or sublinear (logarithm of the term frequency) for each data instance (document). In this study, parameter selected for bag of words model is Term Frequency: counting the number of occurrences of a word in a document. The **Distances** widget calculates distances between rows or columns in a dataset which is normalized to ensure equal treatment of individual features. The resulting distance matrix can be fed further to **Hierarchical Clustering** for uncovering groups in the data. The **Distributions** widget displays the value distribution of discrete or continuous attributes. If the data contains a class variable, distributions may be conditioned on the class. In this study, the graph displayed by the widget shows how many times (e.g., in how many instances) each attribute value appears in the data.
4. Vacancy Analysis with Orange
Following the process describe in section 3, this section presents the results of reliable text mining analysis on vacancy and skills for analyst in Malaysia.

4.1. Dataset Exploration
The results begin with visualization of the most frequent word from 465 document on job title and job description of unstructured data. The top eighteen words with its frequency value are shown in figure 4 and reference job positions (manager, senior, etc), analyst occupational area (business, accounting, financial, operation, support) and technologies (software, data, etc). Analysis from this attribute provides valuable insight into demand for analyst position and skills in Malaysia Labour market in 2019. Figure 5 adds more information from the content of job description analysis. It illustrates the importance of location, technology, responsibility and skill for a description of an analyst vacancy.

Figure 3. Text clustering workflow.

Figure 4. Top terms - Analyst dataset (Job Title).
4.2. Text Clustering Analysis

Orange for data mining supplies processes for computing distance between records in a dataset. Through comparison of textual attribute and similarity ranking construction, this study further analyze the document by identifying potential clusters of job vacancies. It was done by identifying rows potentially belonging to the same category. With similarity ranking from the distance measure, we can identify the closest and farthest records of the data sets that will form several homogenous within and heterogeneous between the cluster. Figure 6 and figure 7 illustrate outcome of hierarchical clustering that visualized the analyst job title into 5 major categories namely C1: Senior analyst, C2: Financial analyst, C3: Business analyst, C4: Accounting analyst and C5: Support analyst.

Figure 5. Top terms - Analyst dataset (Job Description).

Figure 6. Hierarchical Clustering Output.
Table 1 shows the details of specific job titles for each job analyst category. In this instance, we can visualize that the most demanded job categories for analyst is the Support Analyst which also covers Senior analyst, Business analyst and Financial analyst. The least demanded analyst job category is the accounting analyst which only relates to accounting job information.

**Table 1. Detailed Cluster of Analyst Job Title.**

| Cluster 1: Senior Analyst | Cluster 2: Financial Planning Analyst | Cluster 3: Business Analyst | Cluster 4: Accounting Analyst |
|--------------------------|--------------------------------------|----------------------------|-------------------------------|
| Accounts Payable Senior Analyst (China Support) | Financial Planning Analyst (Financial Reporting Accountant)- Priority | | Accounting Sr Analyst |
| Finance & Senior Finance Analyst | Financial Planning Analyst/Accountant | | Accounting Sr Analyst |
| General Accounting Senior Analyst (Indonesia Support) | Financial Planning Analyst (Accountant)- Priority | | Accounts Payable Analyst |
| IT Security Senior Analyst - Regional | Payroll Analyst | | (English/Mandarin/Korean/Vietnamese/Thai) |
| Order-to-Cash (O2C) Senior Analyst (China Support) | Pricing Data Analyst | | Accounts Payable Analyst |
| Order-to-Cash (O2C) Senior Analyst (China Support) | Procurement Process & System Analyst | | (English/Mandarin/Korean/Vietnamese/Thai) |
| Record-to-Report (R2R) Senior Analyst (China Support) | Procurement Vendor Master Data Analyst | | Accounts Payable Analyst 1 (Mandarin Speaking) |
| Senior Analyst, General Accounting (GL & Fixed Assets) | Regional Risk Manager, Data Analyst | | Accounts Payable Analyst |
| Senior Analyst, Transaction Monitoring | Regional Risk Manager, Data Analyst | | (English/Mandarin/Korean/Vietnamese/Thai) |
| Senior Analyst, Transaction Monitoring - NCB, FCSU | Risk Analyst - Market Risk Analytics (Traded) | | Accounts Receivable Analyst |
| Senior Analyst, Transaction Monitoring - NCB, FCSU | Statutory Compliance Delivery Analyst | | (English/Mandarin/Korean/Vietnamese/Thai) |
| Senior Analyst, Transaction Monitoring | Statutory Compliance Delivery Analyst | | Accounts Receivable Analyst 1 (China Support) |
| Senior Finance Analyst | Statutory Compliance Delivery Analyst | | Accounts Receivable Analyst 2 |
| Senior Financial Analyst | Tax Process Improvement Analyst | | (English/Mandarin/Korean/Vietnamese/Thai) |
| Senior II, Business Analyst | | | Analyst Programmer (located in Warehouse, Shah Alam) |
| Senior IT Business Analyst - SAP HRMS | | | Analyst, Account Receivable (Mandarin Speaking) |
| | | | Analyst, Accounts Payable (Japanese Speaking) |
| | | | Analyst, Accounts Payable (Japanese Speaking) |
| | | | Billing Analyst (SAP) - 1 year contract (SSC) |
| | | | Compensation & Benefit Analyst (Portuguese/Spanish) |
| | | | Credit Analyst |
| | | | (English/Mandarin/Thai/Vietnamese/Tagalog/ Bahasa Indonesia) |

**Figure 7. Distribution - Cluster Output.**
4.3. Word Cloud

A word cloud is a graphical representation of word frequency. The word cloud created using Orange software for job description in reference to analyst job categories are shown in the following figure. All documents were passed into an online word-cloud generator and the words are displayed in figure 8 to figure 12. Analysis of unstructured data on job description for business Analyst category reveals that “support”, “sourcing”, “services” and “data” are being emphasized in discussions about Business Analyst vacancy. Words that are frequently used to describe the Accounting analyst job descriptions are “clients”, “account” and “reporting” (figure 9).
Figure 8. Top similarity term of job description for Business Analyst.

Meanwhile, figure 10 displayed the words that are frequently used to describe the Support Analyst job descriptions are “sourcing”, “company”, “service” and “data”. Figure 11 displayed the words that are frequently used to describe the Financial Analyst job descriptions are “reporting”, “client” and “company”. While in the Senior Analyst category (see figure 11), the frequent words are “business”, “financial” and “support”.

Figure 9. Top similarity term of job description for Accounting Analyst.
Figure 10. Top similarity term of job description for Support Analyst.

Figure 11. Top similarity term of job description for Financial Analyst.

Figure 12. Top similarity term of job description for Senior Analyst.
5. Conclusion
This study analyses job market demands particularly analyst job vacancies which provides directly actionable information about vacancy and skills demand. The findings show the classification results and enhance information extraction from job descriptions. By accessing publicly available vacancy data, with web and text mining tools, the study able to extract valuable facts about competences and abilities sought by employers. This approach supports skills research by providing a fact-based alternative for the resource-expensive employer surveying. However, the findings extract mostly on vacancy information and lacking of skills information. The findings provide occupational analysis which is lacking a reflection of current needs, and leads to the gap between skills demand and supply. This limitation is due to the data obtained from online job posts which mostly display the jobs-related information. Although this study only focus on specific job direction, the steps taken could be applied to any type of job vacancies. Future directions in this work include more datasets from several online job websites and also more advanced text pre-processing and cleaning techniques to further improve the classification results. Furthermore, future works also will involves comparing and combining university skills and industry work skills in order to minimize the imbalance of industry-university skills demand. Therefore, the work will be extended by considering predictive model in order to map the university-industry skills and demands from both parties.

References
[1] Piad K C 2018 Determining the dominant attributes of information technology graduates employability prediction using data mining classification techniques J. of Theoretical & Applied Information Technology, 96(12)
[2] Seng L C 2018 Malaysia public universities’ graduate employability policies: an analysis of first degree graduates unemployment and underemployment issues Int. J. of Social Science and Humanities Research ISSN 2348-3164 (online) 6(4) p 480
[3] Mpho M P and Kaelo M 2017 Addressing employability challenges:a framework for improving the employability of graduates in Botswana Int. J. of Adolescence and Youth 22(4) p 455
[4] Zimmermann J, Brodersen K H, Heinimann H R and Buhmann J M 2015 A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance J. of Educational Data Mining 7(3) p 151
[5] Lim H E 2010 Predicting low employability graduates: The case of Universiti Utara Malaysia The Singapore Economic Review 55(03) p 523
[6] Thakar P and Mehta A 2017 A unified model of clustering and classification to improve students’ employability prediction Int. J. of Intelligent Systems and Applications 9(9) p 10
[7] Mishra T, Kumar D and Gupta S 2016 Students' employability prediction model through data mining Int. J. of Applied Engineering Research 11(4) p 2275
[8] Jantawan B and Tsai C F 2013 The application of data mining to build classification model for predicting graduate employment arXiv (Preprint arXiv/1312.7123)
[9] Sapaat M A, Mustapha A, Ahmad J, Chamili K and Muhamad R 2011 A data mining approach to construct graduates employability model in Malaysia Int. J. of New Computer Architectures and their Applications (IJNCAA) 1(4) p 1086
[10] Daud, S., Abidin, N., Mazuin Sapuan, N., & Rajadurai, J. 2011 Enhancing university business curriculum using an importance-performance approach: A case study of the business management faculty of a university in Malaysia. International Journal of Educational Management, 25(6), 545-569.
[11] Debortoli S, Müller O and vom Brocke J 2014 Comparing business intelligence and big data skills: a text mining study using job advertisements Business & Information Systems Engineering 6(5)
[12] Leon L A, Seal K C, Przasnyski Z H and Wiedenman I 2018 Skills and competencies required for jobs in business analytics: a content analysis of job advertisements using text mining Operations and Service Management: Concepts, Methodologies, Tools, and Applications (IGI
Global)

[13] Gao L and Eldin N 2014 Employers’ expectations: A probabilistic text mining model Procedia Engineering 85 p 175

[14] Karakatsanis I, AlKhader W, MacCrory F, Alibasic A, Omar M A, Aung Z and Woon W L 2017 Data mining approach to monitoring the requirements of the job market: A case study Information Systems 65 p 1

[15] Fong L L, Sidhu G K and Fook C Y 2014 Exploring 21st century skills among postgraduates in Malaysia Procedia-Social and Behavioral Sciences 123 p 130

[16] Buntat Y, Jabor M K, Saud M S, Mansor S M S S and Mustaffa N H 2013 Employability skills element’s: difference perspective between teaching staff and employers industrial in Malaysia Procedia-Social and Behavioral Sciences 93 p 1531

[17] Wowczko I 2015 Skills and vacancy analysis with data mining techniques Informatics 2(4) p 31-49 (Multidisciplinary Digital Publishing Institute)

[18] EduSpiral Consultant Services, “Job Outlook, Demand and Salary for Data Scientists and Data Analysts in Malaysia”, Accessed May 12, 2019 [Online]. Available: https://eduspiral.com/2019/04/10/job-outlook-demand-salary-data-scientists-analysts-malaysia/

[19] Lim S M, “STEM jobs in high demand”, Accessed July 18, 2019 [Online]. Available: https://www.focusmalaysia.my/Income/stem-jobs-in-high-demand