Journal of Quality Measurement and Analysis *JQMA* 17(1) 2021, 177-187
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*

TRACKING EMPLOYMENT TRENDS IN MALAYSIA USING TEXT MINING TECHNIQUE

(Mengesan Trend Pekerjaan di Malaysia Menggunakan Teknik Perlombongan Teks)

SYERINA AZLIN MD NASIR* & WAN FAIROS WAN YAACOB

ABSTRACT

The Covid-19 pandemic has changed the world we live in today. In particular, Movement Control Orders (MCOs) that have been deployed nationwide also have an indirect impact on the job creation. With the large number of graduates who have graduated and those who do not have a job will make it even more difficult to get a job. This study attempts to investigate the employment trends during the pandemic in Malaysia by extracting job advertisements randomly from JobStreet website from September to October 2020. A sample of 1050 documents was analysed using text mining technique on two driving factors, job title and location. The results reveal that the highest number of positions offered are managers and the place that offered the most jobs was in Kuala Lumpur followed by Selangor. Further analysis is performed using K-Mediods Clustering to cluster the job titles against the location to illustrate the employment trends in Malaysia, which resulted in similar outcomes.

Keywords: clustering; job vacancy; text mining

ABSTRAK

Pandemik Covid-19 telah mengubah dunia yang kita hidup pada hari ini. Khususnya, Perintah Kawalan Pergerakan (PKP) yang telah dikerahkan di seluruh negara juga memberi kesan tidak langsung kepada penciptaan pekerjaan. Dengan bilangan siswazah baharu yang ramai bermakna mereka yang menganggur akan lebih sukar lagi untuk mendapatkan pekerjaan. Kajian ini cuba menyiasat trend pekerjaan semasa pandemik di Malaysia dengan mengekstrak iklan pekerjaan secara rawak dari laman sesawang JobStreet dari September hingga Oktober 2020. Sampel 1050 dokumen telah dianalisis menggunakan teknik perlombongan teks atas dua faktor, nama jawatan dan lokasi tempat kerja. Keputusan yang didapati mendedahkan bahawa jawatan yang paling banyak ditawarkan adalah pengurus dan tempat yang paling banyak menawarkan pekerjaan adalah Kuala Lumpur diikuti oleh Selangor. Analisis lanjut dilakukan menggunakan Kluster K-Mediods untuk mengelompokkan nama jawatan terhadap lokasi tempat kerja untuk menggambarkan trend pekerjaan di Malaysia, dengan keputusan yang sama.

Kata kunci: pengelompokan; kekosongan jawatan; perlombongan teks

1. Introduction

The number of unemployed graduates is a major concern by many countries, especially during the current situation of the Covid-19 pandemic. In Malaysia, the unemployment rate measures the number of people actively looking for a job as a percentage of the labour force. Leo (2019) reported that every year, over 290,000 Malaysian students graduate from institutions of higher learning. From this value, 1 in 5 fresh graduates remain unemployed after 6 months of graduation, making up to 55% of those who are unemployed, with the majority being degree holders. This can be further observed from the Department of Statistics

Malaysia (DOSM 2020) which stated the increment of unemployment rate of 3.3% in Jun 2019 to 4.7% in August 2020. The rate dropped from 5.3% in May 2020 as the country gradually lifted strict coronavirus lockdown restrictions. The figures show the high unemployment rate faced by Malaysia, especially in the post Covid-19 outbreak which has a severe impact on the country's economy in general and job creation in particular.

Previous works show that retrieving and analysing the content of job advertisements is made possible through text mining by examining job creation from online vacancies (Syerina Azlin *et al.* 2020; Maer-Matei *et al.* 2019). Work by Maer-Matei *et al.* (2019) adopted text mining approach to extract large amount of data on job advertisements to discover main skills and demand of research positions in Europe. Meanwhile, the same approach is used by Espinoza *et al.* (2015) which focuses on specialisations for the Peruvian statistics professional. Study by Syerina Azlin *et al.* (2020) investigates the Malaysian job demands for analysts by categorising job titles and eliciting vacancies and skills from job descriptions. Apart from that, remote work is seen as an innovative form of employment that is perceived as an alternative to the traditional in-office work environment (Kuligowska & Lasek 2013). Due to the Covid-19 pandemic, the job market will most certainly change forever by work from home (WFH) experiment (Buheji & Buheji 2020). Organisations looked to adapt their ways of working in response to the crisis. Working from home could completely transform the job market and could mean truly global competition for every single job role. Hence, job market could become truly globalised where it creates possibilities of working from anywhere when work from home. Thus, this study will explore the job market patterns by extracting data from the popular job advertisement, JobStreet in Malaysia (JobStreet 2020).

The study aims to investigate the kind of jobs created by industries in Malaysia for the post-covid19 era. The insights can highlight the direction of the job market as well as identifying job patterns for better understanding on current employment trends. In this paper, an explanation of text mining approaches and techniques is given. The detailed description of the text mining approach taken in this study is further elaborated. Then follow with the results of the analysis and finally discussed the implications for the future of research.

2. Related Works

Online job recruitments are actively developed, updated and promoted to offer many benefits since it can reach bigger audience, assesible to individuals and thus, making it a more effective method of getting job advertisements noticed. Hence, it provides a platform for massive number of vacancy data that can be used in analysing volumes of potentially useful information. Maer-Matei *et al.* (2019) highlighted the most significant factor to employment in any job position is a good qualification of the applicant and IT skills, especially for research positions in Europe. Meanwhile, Espinoza *et al.* (2015) discovered that experience is the most required aspect for statistician position in Peru with a skill of SQL language. Findings from Wowczko (2015) indicated that experience, skills, and knowledge are the most important as well as the ability to work within a team and business and customer orientation. Specifically, the skills needed for IT administrator are teamwork and good communication whilst for IT engineer are software development and excellent communication. Whereas, Dake (2018) found that the current employment trends in Ghana based on most search jobs are Education, Engineering, Banking, NGO, Healthcare and Accounting. On top of that, Kuligowska and Lasek (2013) found that there are four most recurrent attributes of remote work offers namely contract duration, types of jobs, candidate's features and linguistic indicators. Thus, in the context of modern, fast changing and unstable economy, the potential skills that required for some jobs will directly affect the employment trends.

Text mining is the process of analysing collections of textual materials in order to capture key concepts and themes and uncover hidden relationship and trends without requiring the precise words or terms to express those concepts (Kino *et al.* 2017). Based on the report written by recruiters at the staffing agency, they created frequency list of keywords, hierarchical cluster analysis and co-occurrence network to determine the quality of the candidates. Kuligowska and Lasek (2013) adopted text mining technique using SAS Text Mining 4.2 software within the SAS Enterprise Miner 6.2 environment to control the issue of staff turnover after the job position has been fulfilled. This is carried out by performing text parsing, clustering and concept linking technique on unstructured text collection of location independent job offers. Darabi *et al.* (2018) used Natural Language Toolkit (NLTK) which utilised in Python to extract the keywords and to parse the text into syntactical part-of speech (POS) in text pre-processing stage. The results showed that the top five technical skills are safety and quality, manufacturing, machine and equipment, programming language and software, and engineering drawings whilst top three professional skills needed in engineering are communication skills, reports and customer service. Besides that, Karakatsanis *et al.* (2017) proposed latent semantic indexing (LSI) model to estimate the demand of occupations by job advertisement documents. The result produced the top 10 most demanded occupation in oil and gas industry as well as banking and finance industry among GCC countries, USA and UK. According to Maceli and Burke (2016), R Statistical package is used to analyse job listings in Library and Information Science and then clustered using Ward's agglomerative hierarchical method to calculate frequency, correlation of terms, generate plots and cluster terms for both job titles and descriptions. Hence, these findings show that the methodology used for text mining mostly involves pre-processing, using various appropriate methods.

3. Methods

3.1. *Data Acquisition*

Text mining approach is adopted in this study to examine all positions from the content of job advertisements in Jobstreet.com.my. Two types of information were retrieved specifically on job titles and locations. A total sample of 1050 vacancies was retrieved from data published from the website domain between September and October 2020 using Import.io app. The sample obtained from JobStreet within 2 months is able to offer more than 95% confidence level at 5% margin of error, and the estimated total population of job offer for 2 months is 4000 (According to JobStreet Covid-19 Job Report (JobStreet-Covid-19 2020), average job offers per two weeks is approximately 1000. The sample size is calculated based on Raosoft (2004)). The data were collected three months after movement restrictions being lifted by the Malaysian government.

3.2. *Text Mining Process*

The method of deriving high-quality information from text is text mining, also referred to as text data mining, roughly equivalent to text analytics (Karakatsanis *et al.* 2017; Kumar & Bhatia 2013). RapidMiner Studio is a data science software platform used to analyse the text to find valuable information. The steps involved in the process of text mining are data pre-processing, data processing and data visualisation.

3.2.1 *Data Preprocessing*

The first step in this pre-processing data was to import the data that had been scrapped from Jobstreet.com website into RapidMiner version 9.7 software (RapidMiner 2020). The process involved is shown in Figure 1. In order to pre-process the data, process of document to data operator is used to generate word vectors from string attribute. Term Frequency-Inverse Document Frequency (TF- IDF) is used for word vectors creation. TF-IDF is a statistical measure used to evaluate how important a word is in a collection or corpus. The last step involves the filtering process to match the given condition.

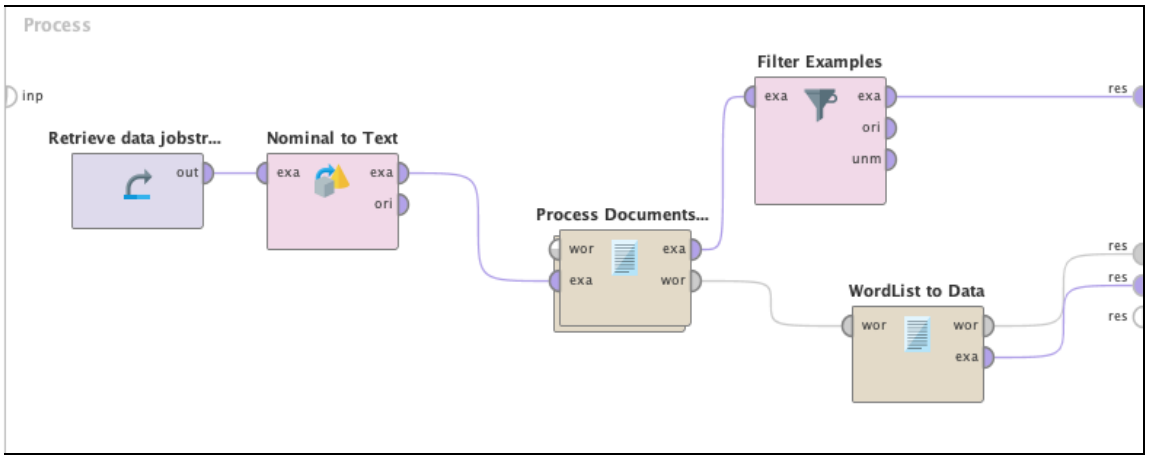

Figure 1: Data Preprocessing Stage

To further clean the data, Process Documents from Data operator is expanded into sub-processes (see Figure 2). In the sub-process, the document is split into a series of tokens using tokenise operator known as 'tokenisation'. Tokenisation will produce several tokens consisting of one single word. Transform Cases operator will transform cases of all characters in a document into lowercase. Filter Tokens (by Length) operator is used to filter the size of the tokens. Filter Stopwords (English) is used to remove English stopwords in a document and replace all occurrences of a specified regular expression to specified replacement using replace tokens operator. The last two sub-processes are Extract Length and Extract Token Number used to extract the document's length and token number before adding them into the metadata to be used later in the analysis.

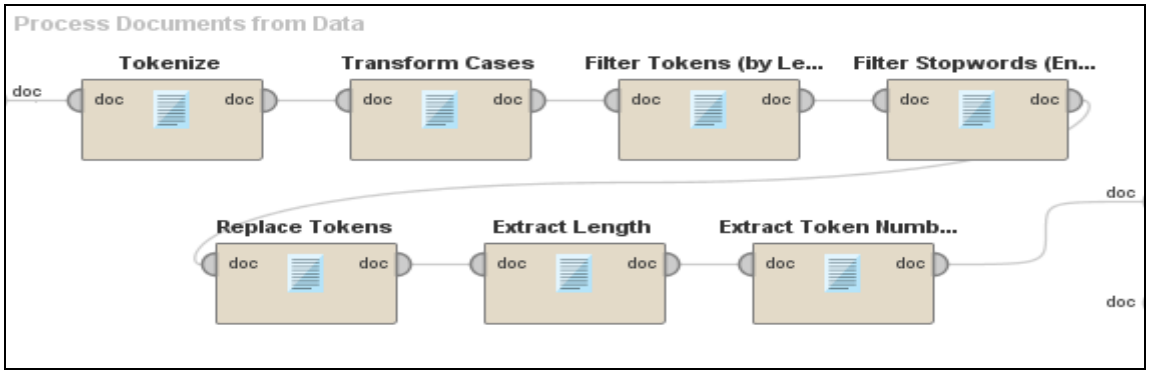

Figure 2: Sub-process of Data Preprocessing Stage

3.2.2. *Data Processing*

Once data is cleaned, the next step in text mining process is to perform data processing. In order to process the data using RapidMiner, Set Role operator is added to the process as illustrated in Figure 3 to change the role of one or more attributes. In this case, to change the text attribute to label target. Store operator is used to store data into the data repository and renamed as 'Job Word Vector'. Worldlist to Data operator is added to convert a wordlist into a dataset and store it in data repository as 'Job WordList'.

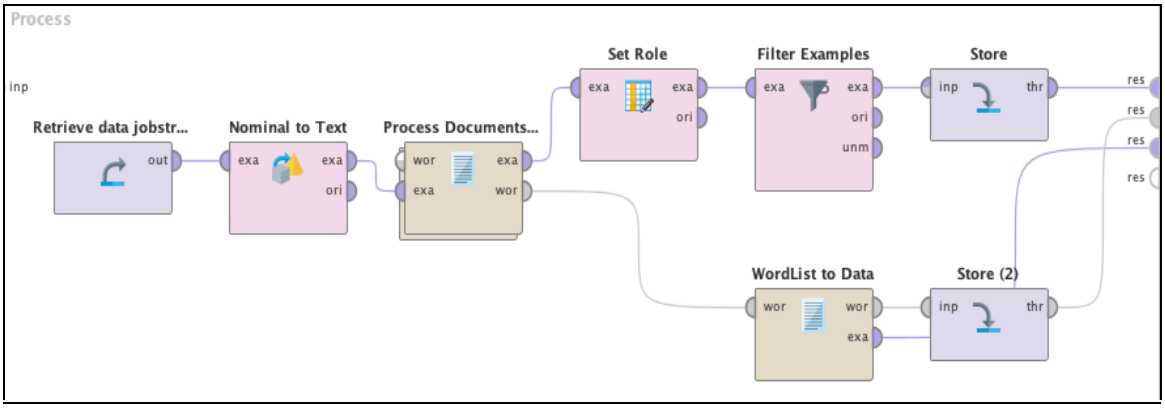

Figure 3: Data Processing Stage

In the Process Documents from Data sub-processes for data processing stage (see Figure 4), Generate *n*-Grams (Terms) operator is added to specify maximal length of n-grams for each token. Term *n*-grams can be defined as a series of consecutive tokens of length, *n*. Then, Aggregate Token Length operator is used to extract and aggregates the length of all tokens of a document and the results will be added as new metadata.

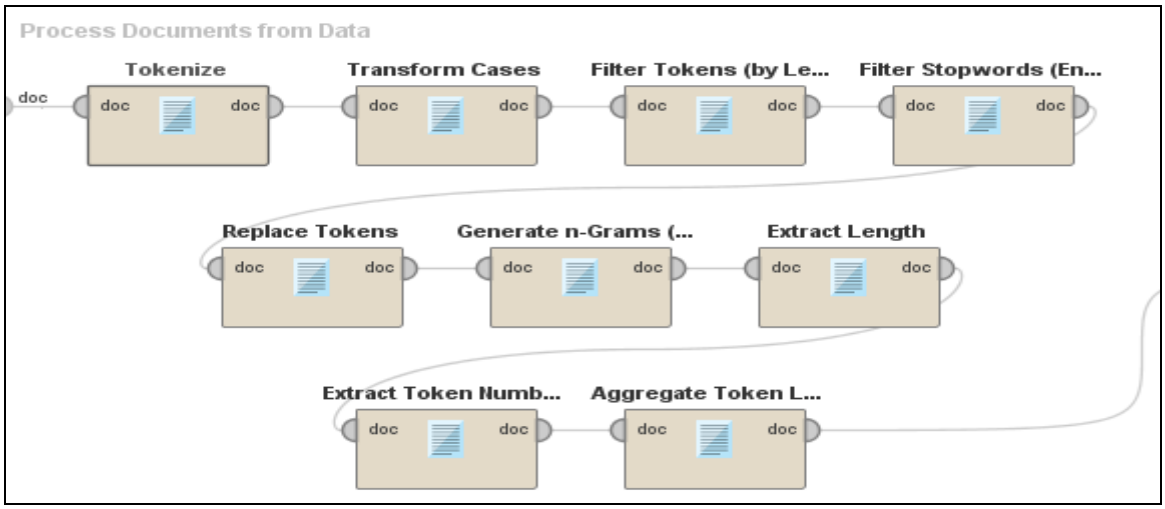

Figure 4: Sub process of Data Processing Stage

3.2.3 *Data Visualisation and Document Clustering*

To visualise the data, the next step in text mining was to generate a word cloud. In this study, the word cloud was produced from the pre-processed data to visualise the word frequency of

the job offers. K-Mediods clustering was also used as the classifier to assign the pre-processed documents into several clusters. K-Mediods is a supervised machine learning method which is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the centre of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with the minimum sum of distances to other points. The scatter plot was conducted to display the clusters of job offers against location to visualise the trend of location and job offer. The classification flow using Rapid Miner is shown in Figure 5.

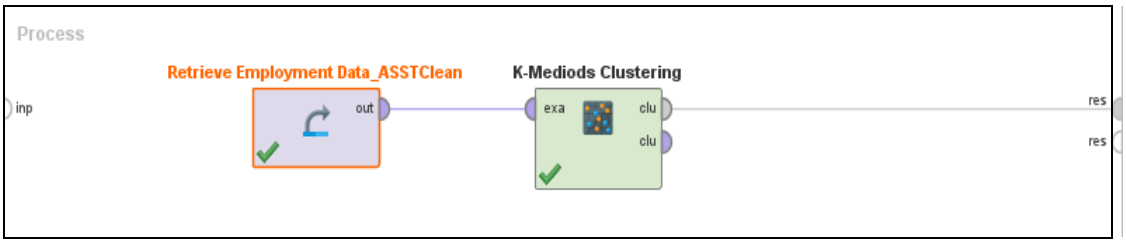

Figure 5: Sub-process of Document Classification

4. Results and Discussion

Figure 6 shows the terms and their frequency in the job lists extracted from the JobStreet posts. Due to the pandemic outbreak, it is to be expected that there will be a decline in the number of job offers. However, the findings from the analysis indicate the opposite, and the most frequent words found for job offers, namely 'manager', 'executive', 'sales', 'assistant' and 'consultant' are still trending even after the second wave of Covid-19 in Malaysia.

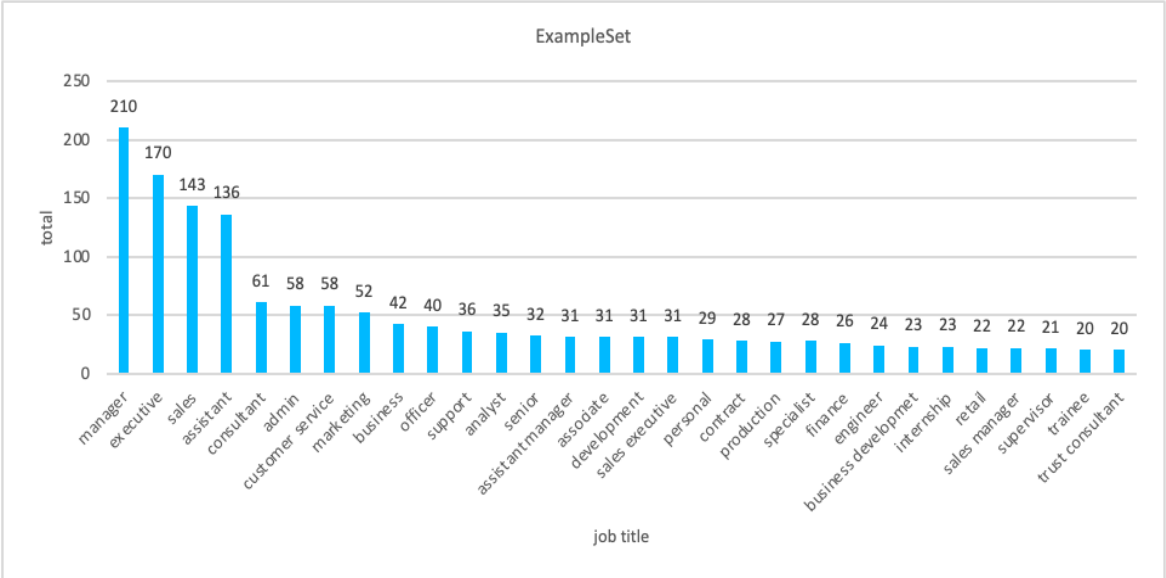

Figure 6: Highest Total Search by Job Title

The results of the job search by location is displayed in Figure 7. The top fifteen terms and their frequency in job lists of job location based on the JobStreet post are displayed below. The top five job location words are 'Kuala Lumpur', 'Selangor', 'Johor Bharu', 'Penang' and

'Petaling Jaya'. This results is also consistent with the finding obtained from the word cloud for job location (Appendix A, Figures A.1-A.3).

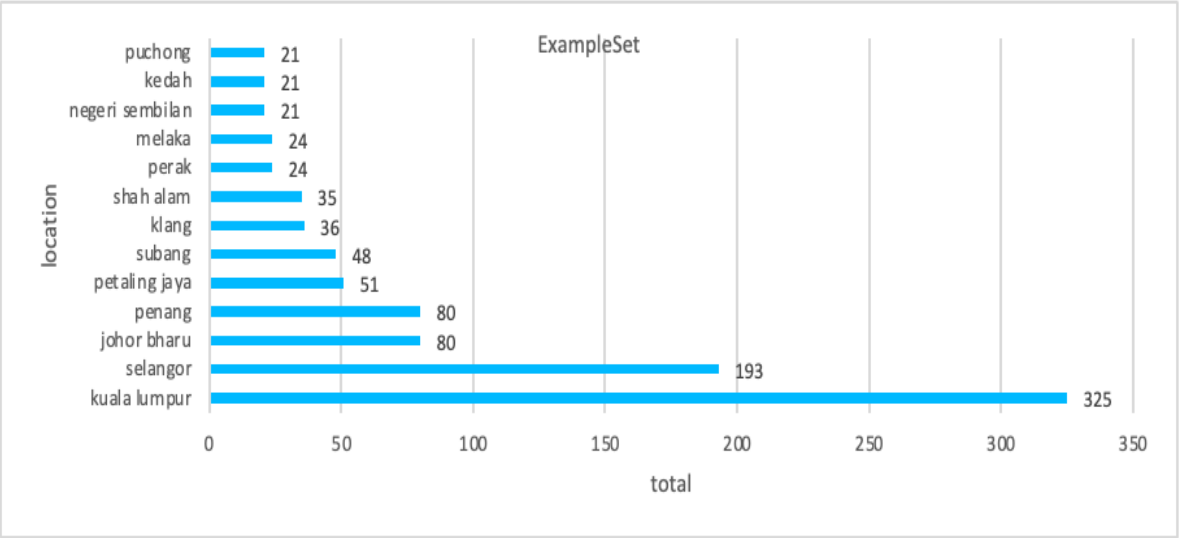

Figure 7: Result of Job Search by Location

A word cloud is a graphical representation of word frequency. The word cloud created using RapidMiner for posts on Job location are shown in Appendix A, Figures A.1-A.3, where A.1 gives the overall result, and A.2 and A.3 are the results on the left and the right sides of the word cloud for readability. The size of each term in the cloud indicates the number of mentions of that term in the posts, reflecting its importance. The dominant word frequency is Kuala Lumpur which indicates high employment opportunities from this location. Thus, analysis from this unstructured data reveals consistent results with Figure 7 where Kuala Lumpur, Selangor, Johor Bahru, Penang and Petaling Jaya are locations that offer the most jobs in the JobStreet.

Based on the finding from K-Mediods Clustering, Figure 8 represents the visualisation of trends for job offers according to a location in each cluster. The scatter plot depicts the irregular distribution of the two attributes in each cluster. Seven clusters of job offers were obtained from K-Mediods Clustering. It can be observed that few of the job clusters display obvious trends clustered at locations in Kuala Lumpur (green dotted points) and Selangor (black dotted points). It shows that these two locations offer various and most jobs during the pandemic.

5. Conclusion

The study mainly utilises large amount of job advertisements for various kinds of jobs ranging from manager, marketing, finance, analyst, engineer, consultant, trainee and many more. Text mining techniques are adopted to further eliminate unwanted information before applying K-Mediods Clustering to further observe Malaysia's employment patterns. Results show that even after the second wave of Covid-19, new jobs appear to still emerge on the market. This is evidenced by the relatively encouraging number of jobs in some places in the rapidly developing states in the country. This study will be continued in the future to study the effects of the third wave of Covid-19 that is currently plaguing the country on the job market.

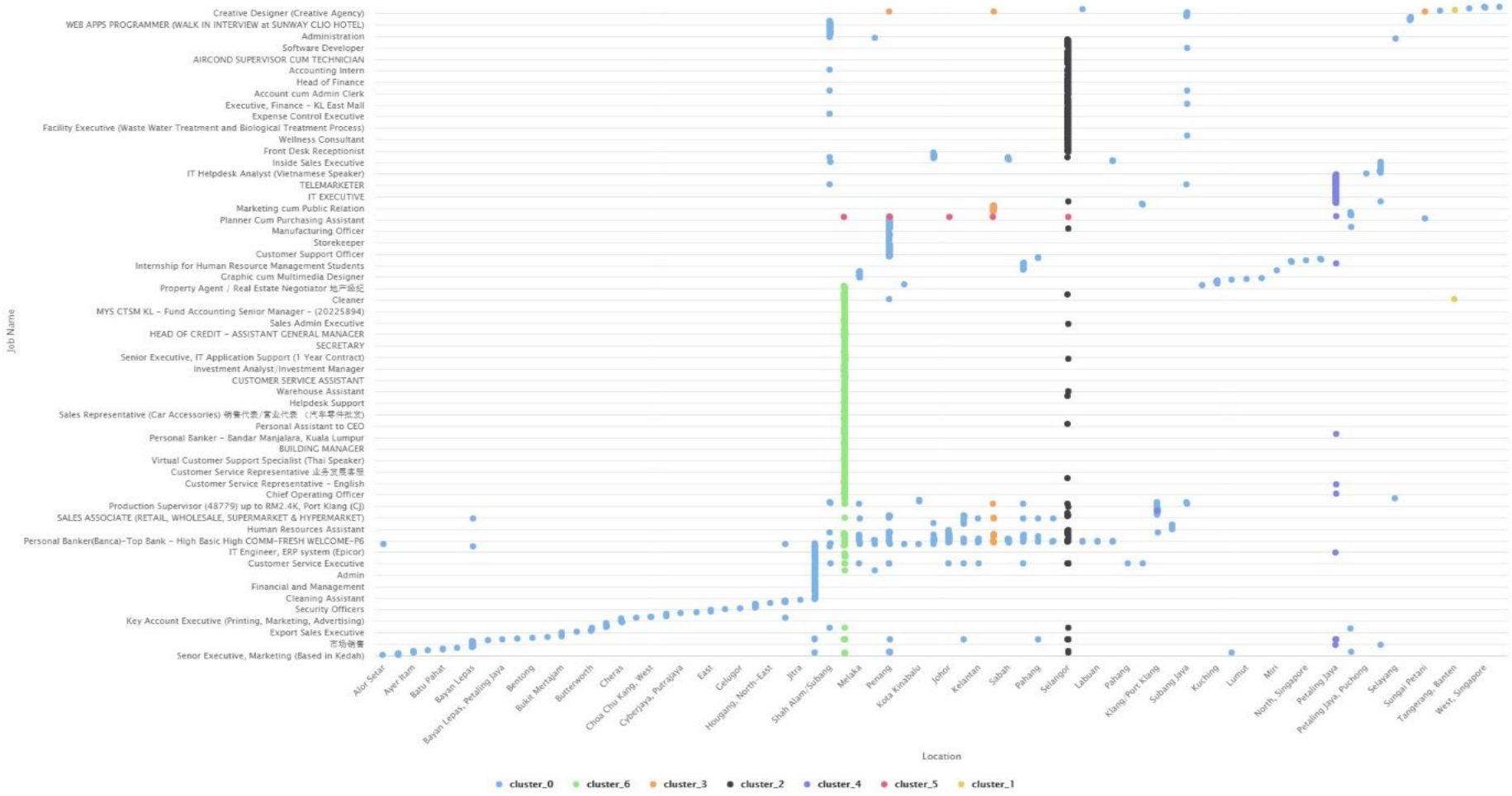

Figure 8: Clustering of Job Offer against Location

References

Buheji M. & Buheji A. 2020. Planning competency in the new normal – employability competency in post-COVID-19 pandemic. *International Journal of Human Resource Studies* **10**(2): 237-251.

Dake D. K. 2018. Using text mining algorithm to track job seeker search patterns in Ghana. *International Journal of Innovative Research in Computer and Communication Engineering* **6**(1):195–201.

Darabi H., Karim F.S.M., Harford S.T., Douzali E. & Nelson P.C. 2018. Detecting current job market skills and requirements through text. Paper presented at the *2018 ASEE Annual Conference & Exposition*, Salt Lake City, Utah. doi:10.18260/1-2--30284.

DOSM. 2020. Department of Statistics Malaysia Official Portal, Gov.my. https://www.dosm.gov.my/v1/index.php (23 Oct 2020).

Espinoza L. C., Guerrero A. R. & Agudo T. N. 2015. Specializations for the Peruvian professional in statistics: a text mining approach. *Proceedings of 2nd Annual International Symposium on Information Management and Big Data*, pp. 35–42.

Karakatsanis I., AlKhader W., MacCrory F., Alibasic A., Omar M. A., Aung Z. & Woon W. L. 2017. Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems* **65**: 1-6.

Kino Y., Kuroki H., Machida T., Furuya N. & Takano K. 2017. Text analysis for job matching quality improvement. *Procedia Computer Science* **112**:1523–1530.

Kuligowska K. & Lasek M. 2013. Text mining in practice: Exploring patterns in text collections of remote work job offers. *Business Informatics* **4**(30): 181–195.

Kumar L. & Bhatia P. K. 2013. Text mining: concepts, process and applications. *Journal of Global Research in Computer Science* **4**(3): 36-39.

JobStreet. 2020. JobStreet Malaysia website, com.my. https://www.jobstreet.com.my (3 Nov 2020).

JobStreet-Covid-19. 2020. JobStreet Covid-19 Job Report, Malaysia August 2020 Edition. https://www.jobstreet.com.my/en/cms/employer/wp-content/themes/jobstreet-employer/assets/loa/report/my/JobStreet-COVID-19-Job-Report-Malaysia-Sept-2020.pdf (23 Oct 2020).

Leo M. 2019. What you didn't know about fresh graduate unemployment in Malaysia. EduAdvisor.my, 26 Aug 2019.

Maceli M. & Burke J. 2016. Technology skills in the workplace: Information professionals' current use. *Information Technology and Libraries* **35**(4): 35–61.

Maer-Matei M. M., Mocanu C., Zamfir A.-M. & Georgescu T. M. 2019. Skill needs for early career researchers - A text mining approach. *Sustainability* **11**(10), art. 2789.

RapidMiner. 2020. RapidMiner website – RapidMiner Studio 9.7 .com. https://docs.rapidminer.com/9.7/studio/releases/changes-9.7.0.html. Released: June 2nd, 2020 (10 June 2020).

Raosoft. 2004. Raosoft Sample Size Calculator. Raosoft Inc., Seattle. http://www.raosoft.com/samplesize.html (24 May 2021).

Syerina Azlin M. N., Wan Fairos W.F. & Wan Adibah Hanis W. Z. 2020. Analysing online vacancy and skills demand using text mining. *Journal of Physics Conference Series* 1496, art. 012011.

Wowczko I. 2015. Skills and vacancy analysis with data mining techniques. *Informatics* **2**(4): 31–49.

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Kelantan
Kampus Machang
Bukit Ilmu, 18500 Machang
Kelantan DN, MALAYSIA
*E-mail: syerina@uitm.edu.my**

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Kelantan, Kampus Kota Bharu
Lembah Sireh, 15050 Kota Bharu,
Kelantan DN, MALAYSIA
E-mail: wnfairos@uitm.edu.my

Received : 16 March 2021
Accepted : 21 July 2021

*Corresponding author

Appendix A

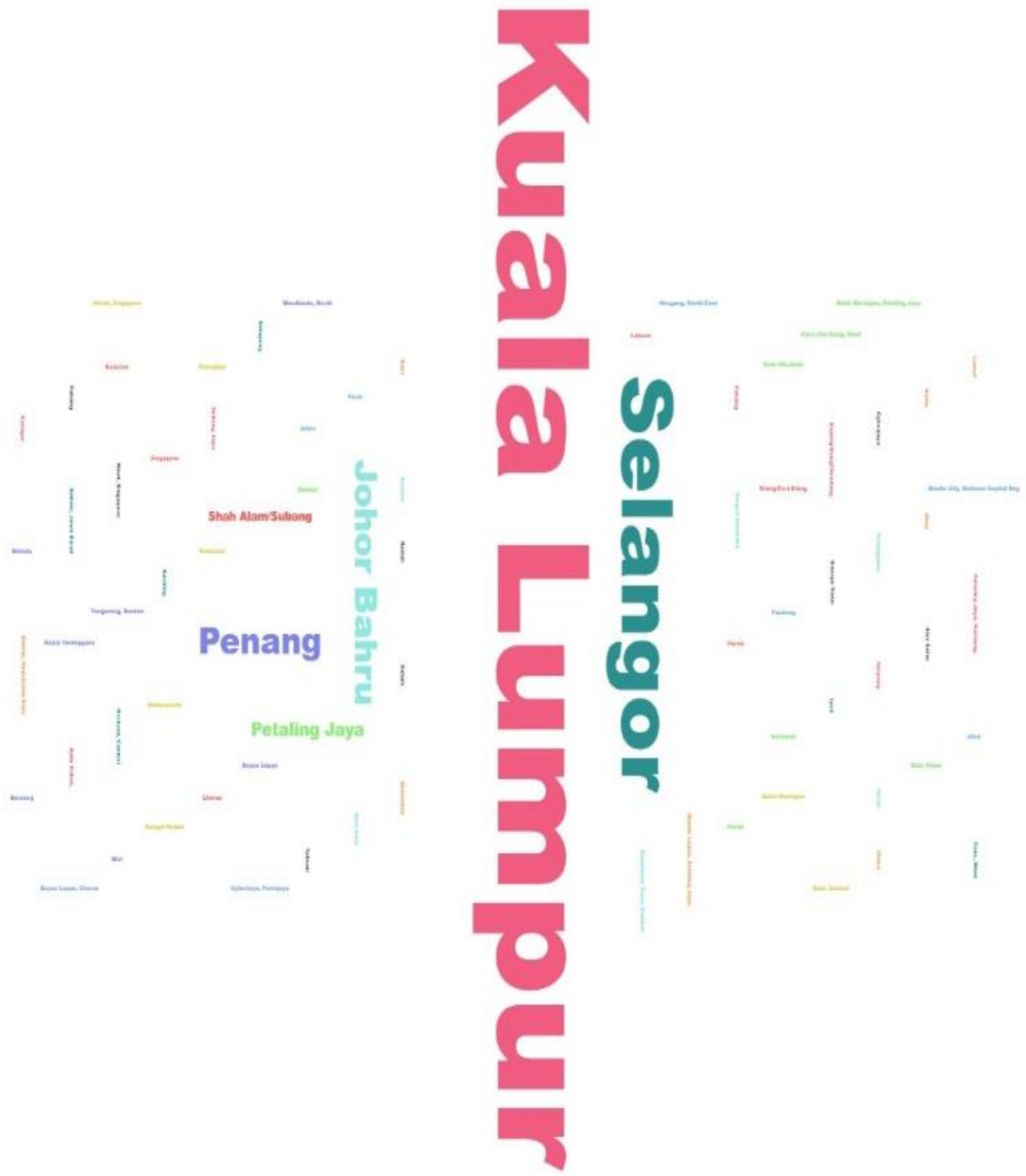

Figure A.1: Word Cloud associated with Job Location (Overall)

Figure A.2: Word Cloud associated with Job Location (Left side)

Figure A.3: Word Cloud associated with Job Location (Right side)