**International Journal of Information Technology and Computer Science Applications (IJITCSA)**
**p-ISSN: 2964-3139 e-ISSN: 2985-5330**
Vol. 02, No. 02, page 75 - 82
Submitted 05/04/2024; Accepted 03/05/2024; Published 10/05/2024

# Harnessing Text and Web Analytics to Enhance Decision-Making in Job Opportunity Categorization

**Santorini Surabani**
*School of Computing, Universiti Utara Malaysia*
*e-mail : santorinisurabani@gmail.com*

**Corresponding Autor:** Santorini Surabani



## Abstract
Text analytics is defined as a method of analyzing compilations of structured text such as dates, times, locations, semi structured text, such as HTML and JSON as well as unstructured text, such as word documents, videos, and images, to extract and discover trends and relationships without requiring the exact words or terms to convey those concepts. Web analytics on the other hand is the technology that collects, measures, analyses, and provides reports of data on how users use websites and web applications. It is used to track a number of aspects of direct user-website interactions, such as the number of visits, time spent on the site, and click pathway. It also aids in the identification of user interest areas and the enhancement of web application features. We used clustering techniques to categorize the job opportunities that are available for the job seekers. By implementing text analytics, text data may be grouped with the goal of providing outcomes in the form of word frequency distribution, pattern identification, and predictive analytics. Text analytics may create one-of-a-kind values to use in the improvement of decision-making and business processes, as well as the development of new business models.

**Keywords**: Text analytics, Clustering, Job Opportunities, Predictive Analytics, Web Analytics

## 1 Introduction

In today's era of globalisation, Big Data is frequently linked to an increase in real-time data acquired from social media, and online portals such as job websites. According to Mezzanzanica and Mercorio [1] big data is a technology that enables businesses to get value from massive volumes of data to better track the adoption of services in the market. In other words, big data is concerned with analytical processes including a mix of data volume, velocity, and diversity, which may involve advanced algorithms and multiple data types. In this research paper, a study on the application of text analytics on online job postings is conducted.

### 1.1 *Data Analytics on Job Postings*

Text analytics is defined as a method of analysing compilations of structured text such as dates, times, locations, semi structured text, such as HTML and JSON as well as unstructured text, such as word documents, videos, and images, to extract and discover trends and relationships without requiring the exact words or terms to convey those concepts [2]. Text mining is described as the process of distilling relevant information from textual insights [2].

Some text mining analyses are semantic and sentimental. Briefly, sentiments represent subjective, human feelings expressed into groups of emotions be it positive, negative or neutral whereas semantics are objective thus deal with understanding the relevant meanings of the data [3]. Such text analyses can be incorporated in analysing job postings in order to explore various relationships as well as trends between keywords from a variety of job titles, skills, reviews and requirements. For instance, sentiment analysis can be implemented through social media analytics. Data extraction and analytics of Social Media Analytics are usually carried out on WhatsApp, Twitter, Facebook, Blogs, and other social networking sites [4]. In most cases, social media sites provide input for sentiment analysis. For example, comments or reviews from the job websites or social media accounts provided by job seekers are crawled to carry out further analysis on them. This way we can tell a job seeker's point of view in terms of whether they are satisfied or not with a particular job scope. Not only that, but geospatial analysis could also be implemented when

performing text analysis such as during the web crawling process where text from job postings' descriptions or job seekers' comments are crawled according to their location or in terms of web analytics where users' location when accessing the website can be accessed and analyzed.

Web analytics on the other hand is the technology that collects, measures, analyses, and provides reports of data on how users use websites and web applications. According to Dange [5] it is used to track a number of aspects of direct user-website interactions, such as the number of visits, time spent on the site, and click pathway. It also aids in the identification of user interest areas and the enhancement of web application features. Websites seek to keep visitors on their pages longer in order to entice them to return and to ensure that each visit ends with the completion of a specific activity. For example, web analytics can be carried out on online job portals in terms of collecting and analysing data of job seekers click pathway in searching for a job of interest or how long does a user usually spend time seeking for jobs on these job portals.

### 1.2 *Literature Review*

Previous research has shown that it is possible to gather insights from online portals using text mining techniques. McGowan [6] incorporated a text mining technique in their project to analyse the search terms for analysts and librarians by categorizing titles and collecting book descriptions. Work by Sinha et al. [7] presented a practical methodology where web scraping is used to collect data from job postings, and text mining is used to extract information about in-demand skills. On the other hand, Goldfarb et al. [8] emphasize job postings related to Artificial Intelligence and Machine Learning to aid curricular development with the type of skills transformation in accordance to the rapid adoption of technologies.

## 2 Problem Statement

As Malaysia enters the rapid globalisation phase, more digital jobs are being offered hence why it is necessary for fresh graduates to be highly skilled to fulfil the present industrial demands. 1 out of every 5 over 290,000 fresh graduates each year remain unemployed after 6 months of graduating, accounting for up to 55% of those unemployed [9]. The Department of Statistics Malaysia claimed that the unemployment rate increased from 3.3% in June 2019 to 4.7% in August 2020 [10]. Therefore, analytics on online postings are highly required to address some of the major problems.

### 2.1 *Skills or labour mismatch*

In research from Bhorat [11] labour mismatch is defined as an inefficient allocation of resources in the labour market between supply and demand. Bhorat [11] even with the online labour market being a major presence that provides valuable information, there still seems to be a gap between real data on qualifications and the skills required by the industry. Fresh graduates on the other hand, are unclear of specific capabilities or job descriptions by certain industries that indicate the skills required hence it is crucial for policymakers and educational bodies to focus their efforts on coming up with a method in bridging the gap between demand and supply for skills.

### 2.2 *Difficulty in decision-making among job seekers*

Due to the unlimited amount of information provided by online job postings from job portals, the decision-making process remains at a complex level. Hence when presented with different options in the online job market, job seekers, particularly fresh graduates, feel overwhelmed and unsure. According to Deming [12] job seekers continue to fall short of the desired level of accuracy and speed in decision-making which is why the usage of computer systems [13], especially with the adoption of data mining algorithms such as text analytics, might aid in the decision-making process. Therefore, the results of text analytics could help fresh graduates in easing the decision-making process. For example, producing charts showing the chosen job title's keywords and skills could help an individual compare the skills they gained from university and the skills needed by the particular job.

## 3 Proposed solution

In this research, text mining algorithms such as clustering will be utilized to gather data of online job postings and job seekers behaviours from job portals as well as social media. The findings are then summarized in the form of data visualization insights and displayed via an analytical dashboard. Particular competencies from job descriptions are captured to identify the skills required for each area of job title available in the current job market according to their location. The required data will be scraped from online job portals and social media, processed, as well as filtered using a set of keywords related to the particular job title to address the problems stated above.

### 3.1 *Data Integration*

Unstructured data cannot be easily merged and examined using a relational database management system (RDBMS). The goal of this research is to bridge the gap between unstructured and organised data by converting unstructured input into structured column values and mapping them to database entities. In this research, we propose an unstructured data integration and analysis system that analyses online job postings using text analytic approaches to extract important information.

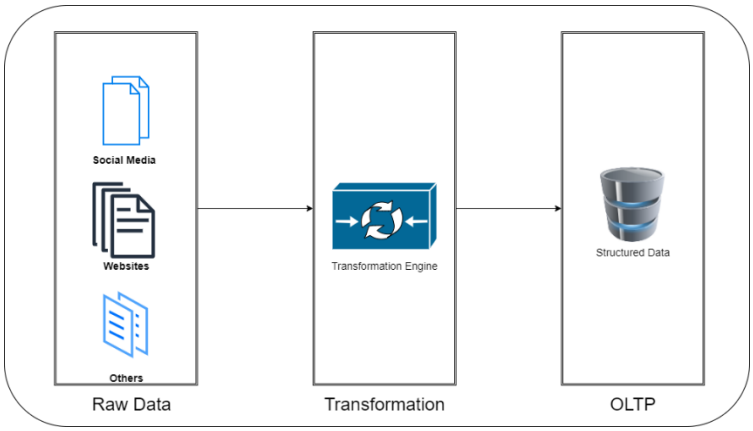


Figure 1. Initial Framework

### 3.2 *Extraction, Transformation, and Load (ETL)*

Change capture, data access, and data federation are three major techniques in data integration that allow data to be obtained and made accessible to a number of ETL and analytic tools as well as the data warehousing environment. Online job postings can benefit from the implementation of a data warehouse that combines data from several sources through the process of ETL.

The ETL process extracts and reads data from one or more sources of databases, such as Facebook, Twitter, LinkedIn, and other social media and job portal platforms, for data collecting. The process of transforming extracted data from one format to another so that it may be placed in a data warehouse is known as transformation. The data is transformed with the use of rules or lookup tables, or by merging it with additional information. The three database functions are merged into a single tool for pulling data from several databases and putting it in a single consolidated database or data warehouse.

### 3.3 *Data Warehouse*

When migrating data to a data warehouse, it is necessary to extract data from all relevant sources for job postings as shown in Figure 2. Data sources include files obtained from online job portal databases, social media tweets, postings and web analytics data. Typically, all of the input files are written to a set of staging tables to make the loading process easier. A data warehouse has a set of business rules that govern how the data will be utilised, as well as summarization, attribute standardisation, and calculation rules. Any data quality problems with the source files must be handled before the data is loaded into the data warehouse.

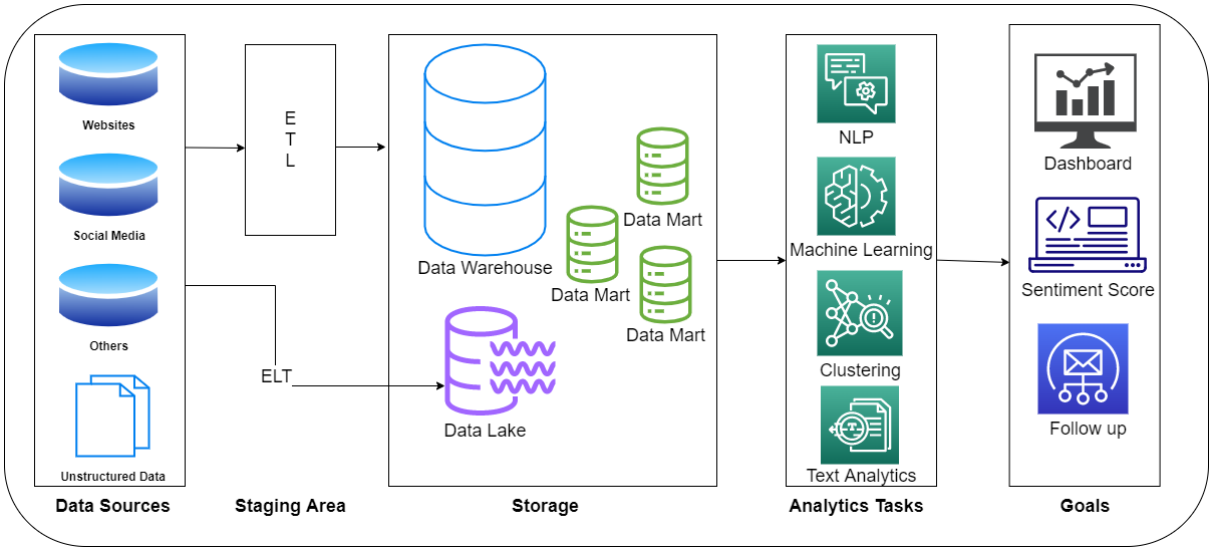


Figure 2. Data Warehouse Infrastructure

## 4 Result and Discussion

To come up with a dashboard as shown in Figure 3 which contains insights on online job postings, this research utilizes Azure cognitive analytics as well as Azure sentiment also known as text analytics. The following visualizations are produced with the help of Power BI to understand the data of job postings, job portals and job seekers better.

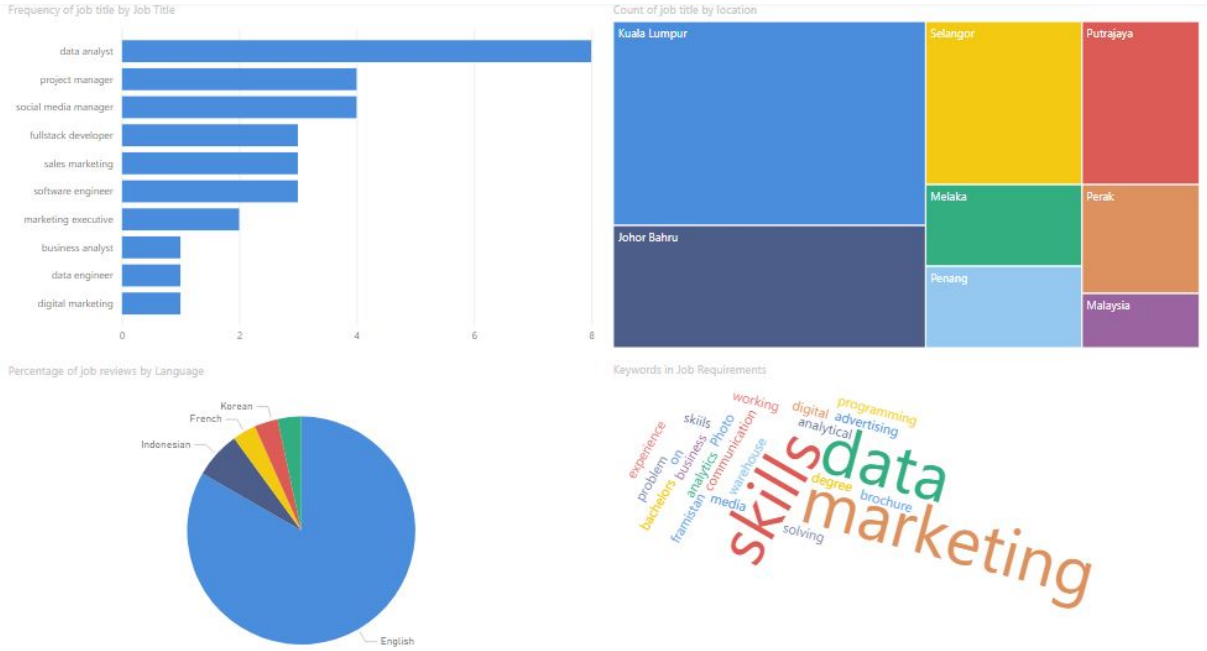


Figure 3. Dashboard on online Job postings

The first visualization represents a frequency bar chart of job titles. From the bar chart, it can be analyzed that the job title with the highest frequency of 8 is “Data Analyst” followed by “project manager” and “social media manager” with a frequency of 4 each. “Business Analyst”, “Data Engineer” and “Digital Marketing” have the lowest frequency each with a value of 1. In terms of the tree map, the job postings’ locations are all located in Malaysia where Kuala Lumpur takes up the highest number of job postings with Johor Bahru staying closely behind.


Figure 4. Pie chart of Job Review Languages

As for the pie chart, job reviews are being analysed in terms of what language they represent. Based on Figure 4 above, the job reviews are mostly in English with a frequency percentage of 83.33%. The other languages available are Korean, Malay, French and Indonesian. The last visualization in the dashboard above is the word cloud. The word cloud represents the keywords associated with job requirements. It can be seen that "skills", "data" and "marketing" are the most popular keywords among job postings.

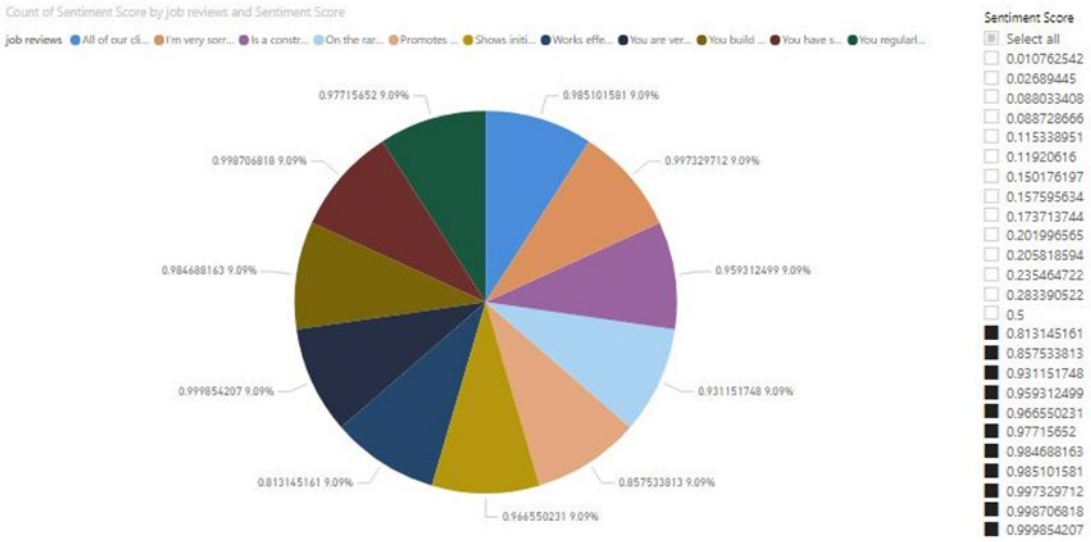


Figure 5. Sentiment Score of Job Reviews

Part of a sentiment analysis is to find the sentiment score of job reviews where the score of above 0.5 to 1 is considered as positive sentiment. Figure 5 above, represents the job reviews with a positive sentiment score. The highest sentiment score is 0.999854207 which is almost a perfect 1. This means that job seekers are highly satisfied with the following job posting, hence the positive review. Figure 6 below represents web analytic insights on an online job portal.

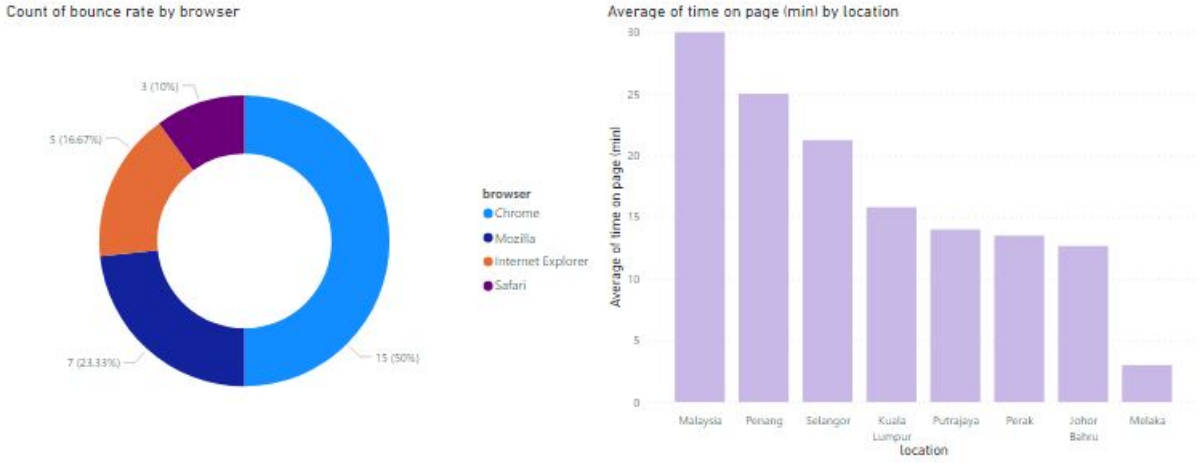


Figure 6. Web Analytics Visualizations

Based on Figure 6, among the four most common web browsers "Chrome" takes the lead with the highest number of counts in terms of bounce rate. The bounce rate is obtained using the formula of dividing the total number of sessions on the site by the number of single-page sessions. As for the average of time spent by job seekers on a page by location, Malaysia takes up the highest average of time with an average of 30 minutes whereas Melaka has the lowest average of time of 3 minutes.


Figure 7. Geospatial Analytics

According to Figure7 above, in terms of geospatial analysis a geo map is used to represent the countries from where job seekers log in to the job website. It can be seen that the job seekers are all from Asian countries. Majority of job seekers log in to the website from Malaysia thus making it the country with the highest frequency of 7 whereas Korea and Thailand are the countries with the lowest frequency of 2.

## 5 Best Practices and Challenges

### 5.1 *Data collection process*

The data collection process starts off with performing web crawling on job portals such as Indeed.com and social media websites such as Twitter and Facebook containing reviews of job postings from job seekers.

### 5.2 *Web Crawling*

Web crawler is a script that crawls the Internet in a systematic and automatic manner. These crawlers are programmes that retrieve web pages and store them in a database. Crawlers create a replica of the recorded web pages, which is subsequently processed by a search engine, which indexes the downloaded pages to aid in rapid searches. In this paper, web crawling is used to crawl text of job seekers' comments on particular job posts or tweets on social media as well as job postings according to location, title, and requirements. Figure 8 below displays a flow chart of the Web Crawler process.

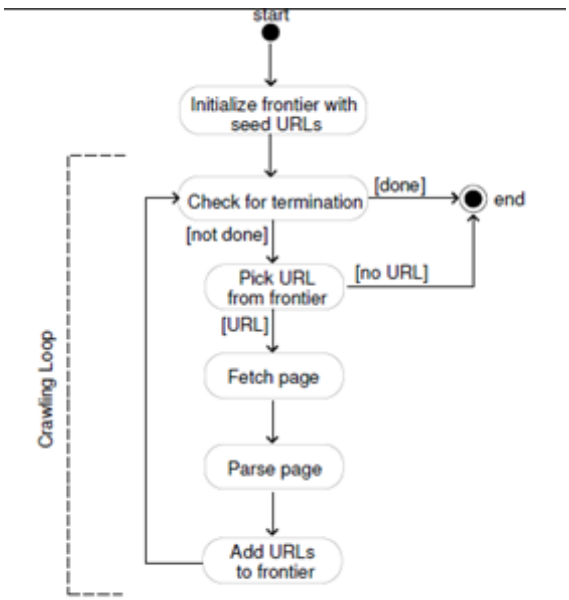


Figure 8. Flow chart of Web Crawler

### 5.3 *Data Cleaning*

A crucial stage in the data analysis process is preparing data as poor data can lead to inaccurate analysis if not addressed. Some approaches that could be used to clean text are normalizing where it is a process of reducing a word's affixes and occasionally derivationally related forms to a single base form. Normalizing techniques include tokenization and removing punctuation. Tokenization is the technique of segmenting flowing text into sentences where text is divided into tokens whereas removing punctuations is a step in removing selected characters from a string.

### 5.4 *Data Mining*

Data mining is implemented in terms of a text mining algorithm such as clustering. Clustering is a process of grouping documents in a large collection that are similar in characteristics into clusters to determine their similarities. Clustering is usually implemented for condensing and pattern recognition in data. The k-means clustering algorithm divides n documents into k-clusters in the context of text data based on the distance between points and cluster centres. The four basic steps of K-Means are as follows:

1. Determining the centers
2. Assigning points to clusters that are outside of the centres based on their distance from in between the centres and points.
3. Calculating the new centers.
4. Repeating steps 1 to 3 till the desired clusters have been obtained.

### 5.5 *Data Visualisation*

As the world accelerates further into the "age of Big Data", data visualization becomes a significant tool for making sense of the unlimited number of rows of data created daily especially with job postings. For example, a word cloud organises keywords by word frequency, then arranges them according to defined rules and visualises them with graphic attributions such font size and colour. Due to its readability, understandability, and simplicity, word clouds are the most commonly utilised technique when it comes to determining the current trends in keywords from job descriptions.

## 6 Challenges and Implementations

A main challenge of implementing the K-means clustering algorithm as it is sensitive to determine the initial points of the number of k. If a poor decision is made at the beginning, many changes will occur over the clustering period, thus different clustering outcomes may be produced with the same number of iterations each time. Besides that, the k-means method involves dealing with outliers. Therefore, implementing k-means itself would be insufficient to reflect the broad set of skill criteria seen in job postings. As for word clouds, they tend to be full of blind spots due to the visibility of words, especially for shorter words, and thus the value assigned to text is commonly imprecise.

## 7 Conclusion

In conclusion, by implementing text analytics, text data may be grouped with the goal of providing outcomes in the form of word frequency distribution, pattern identification, and predictive analytics. Text analytics may create one-of-a-kind values to use in the improvement of decision-making and business processes, as well as the development of new business models. As for the dashboard, with the aid of visual trends, one may quickly and easily determine what the ideal next step in making a decision is in a short amount of time as it simplifies the data and makes it more shareable and available to access.

BIBLIOGRAPHY

[1]. M. Mezzanzanica and F. Mercorio, "Big Data enables Labor Market Intelligence," Encyclopedia of Big Data Technologies, pp. 226–236, 2019. doi:10.1007/978-3-319-77525-8_276.
[2]. B. Tucker, E. Santhanam, and E. Zaitseva, "Future directions and challenges in text analytics," Analysing Student Feedback in Higher Education, pp. 205–217, Dec. 2021. doi:10.4324/9781003138785-18.
[3]. T.-S. Nguyen, Z. Wu, and D. C. Ong, "Attention uncovers task-relevant semantics in emotional narrative understanding," Knowledge-Based Systems, vol. 226, p. 107162, Aug. 2021. doi:10.1016/j.knosys.2021.107162.
[4]. S. Naeemi, "Social Media Actions Analytics," Social Media Analytics in Predicting Consumer Behavior, pp. 111–129, Mar. 2023. doi:10.1201/9781003200154-6.
[5]. A. S. Dange and Dr. M. E, "Text matching technique based intelligent web crawler in hybrid mode," SSRN Electronic Journal, 2022. doi:10.2139/ssrn.4053442.
[6]. B. S. McGowan, "Using text mining tools to inform search term generation: An introduction for librarians," portal: Libraries and the Academy, vol. 21, no. 3, pp. 603–618, 2021. doi:10.1353/pla.2021.0032.
[7]. K. Sinha, P. Sharma, H. Sharma, and K. Asawa, "Web scraping and job recommender system," 2023 Second International Conference on Informatics (ICI), Nov. 2023. doi:10.1109/ici60088.2023.10420941.
[8]. A. Goldfarb, B. Taska, and F. Teodoridis, Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings, Feb. 2022. doi:10.3386/w29767.
[9]. L. M, "Fresh graduate unemployment in Malaysia," EduAdvisor, https://eduadvisor.my/articles/what-didnt-know-fresh-graduate-unemployment-malaysia-infographic (accessed Mar. 5, 2024).
[10]. Dosm, Department of Statistics Malaysia, https://www.dosm.gov.my/v1/index.php (accessed Mar. 5, 2024).
[11]. H. Bhorat, "Links between education and the Labour Market: Narrowing the mismatch between demand and supply," Skill Formation and Globalization, pp. 145–160, Jun. 2019. doi:10.4324/9781351149006-9.
[12]. D. Deming, The growing importance of decision-making on the job, Apr. 2021. doi:10.3386/w28733.
[13]. H. Surbakti, "Pemodelan Arsitektur Enterprise pada Perguruan Tinggi Untuk Peningkatan Layanan Pendidikan (Studi Kasus: Universitas Respati Yogyakarta)," UAJY E-Print Thesis. Jan. 2018. https://e-journal.uajy.ac.id/13582/.