ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

Indonesia Job Vacancy Dataset

[1] Mario E. S. Simaremare*, [2] Triagus A. Simanjuntak, [3] Cintya Y. S. Nainggolan, [4] Imelda S. D. Siregar

[1] [2] [3] [4] Institut Teknologi Del, Laguboti, Indonesia
Corresponding Author Email: [1] mario@del.ac.id, [2] triagusabdi381@gmail.com, [3] cintyaysn277@gmail.com, [4] imeldasiregar02@gmail.com

***Abstract*— Job vacancies are jobs (positions) for employers who are looking for suitable employees (to fill vacancies). Job vacancy data serve as a valuable source of information on the characteristics of labor market demand. For example, information that is often found in job vacancy data is a job title, job description, and qualification. Job vacancies could also be used as an indicator of how good a country's economy is. Until now there is still no research that produces job vacancy datasets in Indonesia. This job vacancy data can be useful to assist other research that requires job vacancy data, such as creating a job recommendation system, identifying trends, and making future predictions.**

In this paper, we would like to share our job vacancy dataset collected from several popular platforms in Indonesia. The platforms we chose to create the datasets were glints.com, jobindo.com, jobstreet.co.id, and karir.com. We developed scrapers, one for each platform, and ran them to grab vacancies posted on the platforms. Afterward, we carefully analyzed the collected vacancy records and designed the pre-processing steps carefully. The steps were data cleaning, transformation, reduction, and integration. In the end, we integrated all the records into one dataset. The final dataset contains 63,870 records with 20 attributes dated from 2015-2022.

***Index Terms*— Data Mining, Dataset, Indonesia, Job Vacancy.**

I. INTRODUCTION

Data is the most important and valuable asset. We can use various data mining techniques, statistical analysis, and machine learning to process and transform them into useful information and actionable knowledge for better decisions in various fields [1]. Commonly, we utilize information and knowledge to predict, compare, recommend, plan, cluster, classify, or identify future trends.

After being struck by a global pandemic, we are now facing a threat of global recession, where the economic sector is predicted to decline [2]–[4]. We are interested in finding an empirical signal indicating the recession's coming. One possible indication is job availability here in Indonesia. After failing to find any current dataset related to job availability, we decided to put some effort into curating records of job vacancies posted on four major platforms used in Indonesia, glint.com, jobindo.com, jobstreet.co.id and karir.com. Job vacancy is useful in describing the requirements to other details related to the job the position will be filled with.

We carefully designed and ran a set of pre-processing steps to make the dataset ready for use. These steps must be able to preserve both its value and its meaning. The pre-processing steps were cleaning, transforming, reducing, and integrating. We will discuss each step further.

In this paper, we share the dataset with other researchers who might want to work in an area where job vacancy or availability is important. Job vacancy data can be used as data for recommendation systems such as recommending the most qualified candidate for a particular job to recruiters or recommending jobs to potential candidates according to their suitable profile [5]. In addition, data can also be used for forecasting, displaying data trends, and demographic data. The availability of such datasets has been shown to be valuable in previous research, such as the datasets provided by [1] and [6] for IT job vacancies in Iran and the analysis of labor market concentration in the United States, respectively. Therefore, our research aims to provide a similar resource for job vacancies in Indonesia.

II. METHODOLOGY

The following section describes our method from platform selection to the description of the final vacancy dataset. On the platform selection process, we chose four platform based on two criteria. We developed scrappers to collect the raw datasets from the platforms. Afterwards, we analyzed and designed a set of pre-processing steps (Fig. 1).

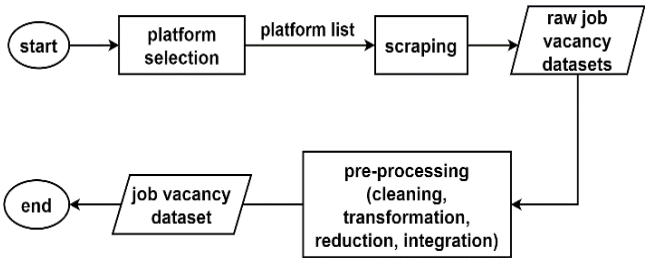

Fig. 1 Research method

A. Site Selection

Our first step was listing all platforms with a job posting feature. Seven platforms were identified (Table I). We employed two selection criteria as follows:

- C1. the platform is still actively used by recruiters to post vacancies and by the job seeker to apply; and
- C2. the platform must allow web scrapping.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

Only the first four platforms passed our selection criteria, they were glints.com, jobindo.com, jobstreet.co.id and karir.com; the other had authorization mechanisms making web scrapping impossible.

Table I: Platform selection

No.	Platform	C1	C2
1.	glints.com	✓	✓
2.	jobindo.com	✓	✓
3.	jobstreet.co.id	✓	✓
4.	karir.com	✓	✓
5.	indeed.com	✓	X
6.	jobsinjakarta.com	X	X
7.	linkedin.com	✓	X

B. Data Collection (Scraping)

The second step was developing a scrapper for each platform due to differences in the web page structures. The scrapers were written in Python. A scrapper would 'walk' through the web pages containing vacancies posted on them. The steps taken are:

a. Create a scraping model
b. Exploring the website by creating a website navigation
c. Automate navigation and extraction
d. Data extraction and package history

Once the page was loaded (Fig. 2) the scrapper parsed it, extracted the vacancies from the page, and recorded them (Table II).

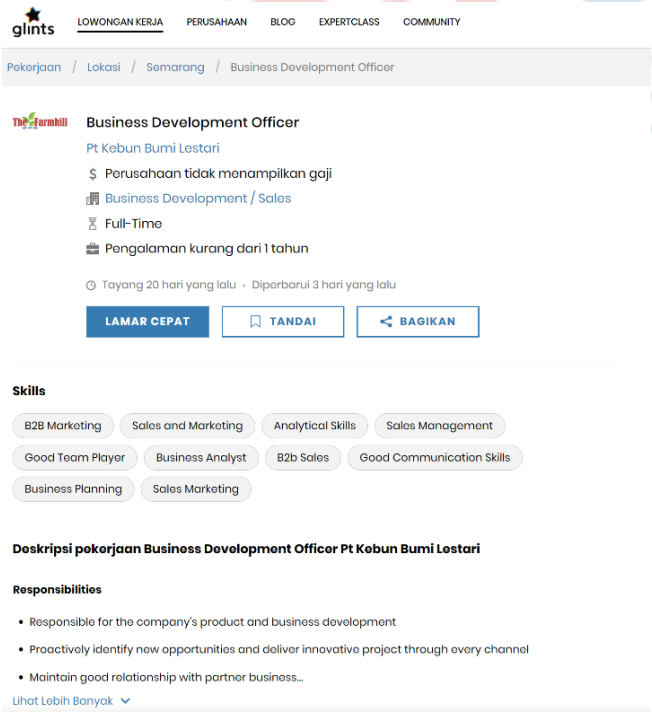

Fig. 2 Job vacancy posted on Glints.com

Table II: Attribute list from job vacancy on Glints.com

No	Attribute	Type	Example
1.	id	nominal	0
2.	logo	nominal	https://images.glints.com/unsafe/glints.com-dashboard.s3.amazonaws.com/
3.	title	nominal	Business Development Officer
4.	company	nominal	Pt Kebun Bumi Lestari
5.	salary	ordinal	Perusahaan tidak menampilkan gaji
6.	time	interval	
7.	skills	nominal	Skills B2B Marketing Sales and Marketing Analytical Skills Sales Management Good …
8.	location	nominal	Semarang
9.	work_ experience	ordinal	Pengalaman kurang dari 1 tahun
10.	type_of_ work	nominal	Full-Times
11.	specializati on	nominal	Business Development / Sales
12.	description	nominal	Deskripsi pekerjaan Business Development Officer Pt Kebun Bumi Lestari Responsibilities Responsible
13.	link	nominal	https://glints.com/id/opportunities/jobs/business-development-officer/0b5ea156-f07b-4c72-b819-...

We stored the recorded vacancies in a plaintext file in CSV format. We called it a raw dataset. In total, we successfully collected 63,870 from the four selected platforms. Table III summarizes the number of vacancies posted on the selected platforms. We gave an id for each platform to ease the future record identification.

Table III: Curated vacancies from the selected platforms

ID	Platform	Starts from	∑ Vacancies
P1	glints.com	-	8,670
P2	jobindo.com	01/12/2015	34,473
P3	jobstreet.co.id	08/12/2022	19,721
P4	karir.com	01/12/2022	1,006
Total			63,870

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

We identified 19 distinct attributes taken from the four selected platforms. Table IV depicts the attribute mapping between platforms.

Table IV: The attribute list of job vacancies from the selected platforms

No	Attribute	P1	P2	P3	P4
1.	id	✓	✓	✓	✓
2.	logo	✓	✓	✓	✓
3.	title	✓	✓	✓	✓
4.	company	✓	✓	✓	✓
5.	location	✓	✓	✓	✓
6.	company_size			✓	
7.	type_of_work	✓		✓	✓
8.	requirement		✓		✓
9.	skills	✓			
10.	specialization	✓			
11.	job_level			✓	✓
12.	total_registrant		✓		
13.	description	✓	✓	✓	✓
14.	work_experience	✓	✓	✓	✓
15.	qualification			✓	✓
16.	work_function				✓
17.	salary	✓		✓	✓
18.	time		✓	✓	✓
19.	link	✓	✓	✓	✓

1) Attribute Selection

These attributes would later be fed into the pre-processing steps. The selected attributes from the platforms are shown in Table IV. The id attribute was not selected because it contained values meaningful only for the specific platform.

Table V: Raw attributes

No	Attributes
1.	logo
2.	title
3.	company
4.	location
5.	company_size
6.	type_of_work
7.	requirement

No	Attributes
8.	skills
9.	specialization

No	Attributes
10.	job_level
11.	total_registrant
12.	description
13.	work_experience
14.	qualification
15.	work_function
16.	salary

No	Attributes
17.	time
18.	link

C. Data Pre-processing

Data gathered in data sets present multiple forms and come from many different sources. Data directly extracted from relational databases or obtained from the real world is completely raw: it has not been transformed, cleansed or changed at all. Therefore, it may contain errors due to wrong data entry procedures or missing data, or inconsistencies due to ill-handled merging data processes. Three elements define data quality: accuracy, completeness and consistency. Unfortunately real-world data sets often present the opposite conditions, and the reasons may vary as mentioned above. There are several preprocessing techniques or steps devised to overcome the problem present in such real-world data sets and to obtain a final, reliable and accurate data [7].

Data pre-processing constitutes of four steps: data cleaning, transformation, reduction, and integration (Fig. 3). The cleaning phase is used to remove noise and inconsistent values. In data transformation, data could be transformed and consolidated into appropriate forms for future processes. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance include process data aggregation, attributes subset selection, dimensionality reduction and numerosity reduction [7]. The last one, data integration, combines data from multiple sources into one [3]. On this research there are no used of data reduction.

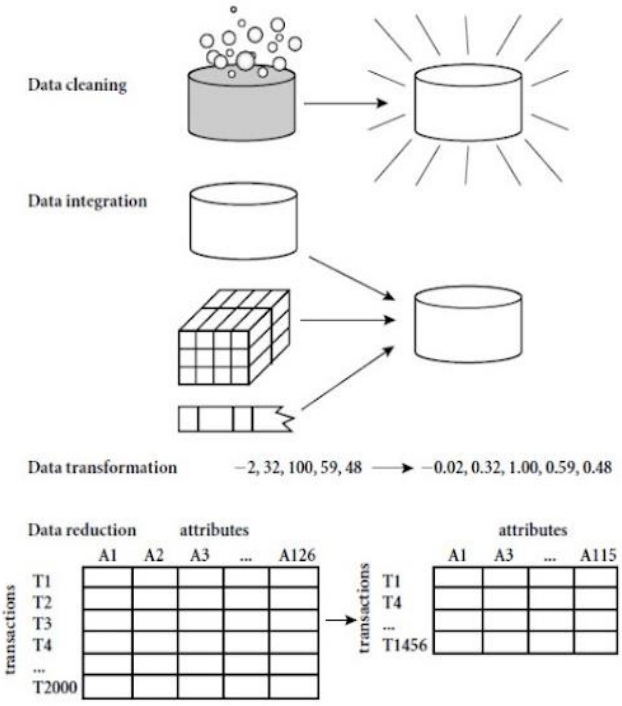

Fig. 3 Methods in data pre-processing [8]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

1) Data Cleaning and Transformation

Data cleaning is useful for filling in missing values, removing noise from outliers, and fixing inconsistent data. Data transformation transforms data into a format suitable for use in data mining. Data transformation can be done through smoothing, aggregation, data generalization, normalization, and attribute construction [8].

We found that some records suffered from missing or empty values (Table VI). These missing values might impact the performance of further processing. Hence, we needed to overcome the challenge. There were two possible strategies available to answer the challenge. The first one was to remove the 'defective' records, or the second option was to fill in the missing values [8]. If we removed data with missing value then useful data from dataset would also be deleted. Therefore, we chose the latter option. We conducted further analysis to find out what values should be used as the replacement.

Table VI: Attributes that were suffering from missing values

P1	salary
	time
	work_experience
	specialization
P2	location
	requirement
	description
	time

P3	logo
	location
	salary
	job_level
	work_experience
	qualification
	type_of_work
	company_size
	description
P4	job_level
	qualification
	description

There are several ways to handle missing values, ignore tuples, fill missing values manually, use global constants, use the mean value of an attribute, use the mean attribute for all samples that belong to the same class as the tuple, and use the same most probable value.

Based on our understanding of the context and by considering the attribute type, we choose to use global constants, here was our strategy to fill in the missing values:

1. Any ordinal or ratio attributes without value were set to '0'.
2. Any nominal attributes without value were set with 'none'. This was applied to the 'job_level' attribute.
3. Minimum or maximum attribute will be filled with 'null' value. This approach was applied to the 'min_work_experience', 'max_work_experience', 'min_salary', and 'max_salary'.
4. Any interval attributes without value were set to epoch time (01/01/1970 00:00:00).

Any attributes that are used but only exist on one platform, it will be equated with adding these attributes to become new attributes on other platform. Some of the attributes constructed on other platforms shown in Table VII.

Table VII: Attributes that only exist on one or several platforms

No.	Attributes
1.	salary
2.	time
3.	specialization
4.	requirement
5.	time
6.	salary
7.	job_level
8.	qualification
9.	type_of_work
10.	company_size

Furthermore, we also found inconsistent values in the datasets. Some examples of this case are shown Table VIII.

Table VIII: Example of inconsistent value

Attribute	Example
work_experience	1 - 3 tahun pengalaman
	Pengalaman 0-2 Tahun
	1 tahun
experience	Diploma 3 (Pengalaman 0-2 Tahun)
	SLTA/Sederajat (Pengalaman 2-5 Tahun)
	Diploma 3 (Pengalaman 2-5 Tahun)
	Sarjana (Pengalaman 2-5 Tahun)
time	6/1/2023 11:59:00 PM
	23/12/22 10:40
	2023-01-12T07:05:30.000Z
salary	IDR 13.000.000 - 20.000.000/ Bulan
	IDR 4M - 5.600.000 per bulan
	IDR 8.000.000 - 10.000.000

Table IX: Consistent value

Old Attribute	New Attribute	Example
work_experience	min_work_ experience	0
	max_work_ experience	1

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

experience	qualification	Diploma 3
	min_work_ experience	0
	max_work_ experience	2
time	time	6/1/2023 11:59:00 PM
salary	currency	IDR
	min_salary	13.000.000
	max_salary	20.000.000

Let's discuss the 'work_experience' attribute as an example. Initially, the attribute stored two possible combinations: the minimum-maximum year of experience or just the minimum year of experience. The earlier combination looked uncommon because of the maximum limit. We chose to preserve the data and let the researchers decide whether to use it. Similar situations also happened to the 'time', and 'salary' attributes (Table IX). Our strategy to handle the inconsistent values is described as follows:

1. The data type of the 'work_experience' attribute is ratio or data that has a range, for example, '1-3 years'. To maximize data processing, every value of the 'work_experience' were transformed into an ordinal value. This was possible by taking the attribute's minimum (min) and maximum (max) values. For example, '1-3 years' was split into two attributes, the 'min_work_experience' attribute with value '1' and the 'max_work_experience' attribute with value '3'.
2. The 'experience' attribute, from P2, had more than one value. The first value was closer to the 'qualification' attribute, and the second was the actual 'work experience'. To handle this issue, we extracted the 'qualification' from it and distributed the rest into two attributes, namely the 'min_work_experience' attribute and the 'max_work_experience' attribute.
3. Similar approach was also applied to the 'salary' attribute, we split the value and introduced two new attributes namely 'min_salary' and 'max_salary'.
4. Standardizing the format for 'time' attribute with 'DD/MM/YY HH:mm'.

2) Data Integration

This was the last step in our method. In this step, we were supposed to merge all four datasets based on the semantics of the attribute values. Hence, we need to map the attributes before merging the datasets. For instance, the dataset retrieved from P1 had 'skills' and 'specification' attributes (Table X) that had the same values as the 'requirements' attribute in another dataset (from another platform). Our approach was to merge the 'skills' and 'specification' attributes into the 'requirement' attribute (Table XI).

Table X: Values from the 'skills' and the 'specialization' attributes (P1 dataset)

'skills' attributes	'specialization' attributes
Skills Persuasive Communication Passionate about Beauty Content Creator	Marketing
Skills Komunikatif Ramah Berorientasi Pelayanan Berkomunikasi Dengan Baik	Other
Skills Marketing Public Speaking Computer	Marketing
Skills Quick Learner Fast Learner Responsibility Creative Thinking Analytical Skills	Business Development / Sales

Table XI: Merge the 'skills' and the 'specialization' attributes into the 'requirement' attribute

'requirement' attributes
Marketing | Skills Persuasive Communication Passionate about Beauty Content Creator
Other | Skills Komunikatif Ramah Berorientasi Pelayanan Berkomunikasi Dengan Baik
Marketing | Skills Marketing Public Speaking Computer
Business Development / Sales | Skills Quick Learner Fast Learner Responsibility Creative Thinking Analytical Skills

III. RESULT

After all of the data pre-processing, we got the complete dataset with 20 attributes. The attributes are listed in Table XII. The total number of records in the dataset is 63,870 and accessible through our GitHub repository[1].

Table XII: Dataset attributes

No	Original Attribute	Final Attributes
1	logo	logo
2	title	title
3	company	company
4	location	location
5	company_size	min_company_size
		max_company_size
6	type_of_work	type_of_work
7	requirement	requirement
8	skills	removed
9	specialization	removed

[1] https://github.com/simaremare/indonesia-job-vacancy-dataset

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

No	Original Attribute	Final Attributes
10	job_level	job_level
11	total_registrant	total_registrant
12	description	description
13	work_experience	min_work_experience
		max_work_experience
14	qualification	qualification
15	work_function	work_function
16	salary	min_salary
		max_salary
		currency
17	time	time
18	link	link

Fig. 4, draws the platforms' record contribution to the dataset. Most of the records were retrieved from jobindo.com (54.0%), followed by jobstreet.co.id (30,9%), glints.com (13.6%), and karir.com (1.6%). This is another view of **Error! Reference source not found.**, served earlier in this paper.

Fig. 5, depicts the job posting trends from 2015 – 2022. From the trends, we can clearly see a decline of job posting during the COVID-19 outbreak (2019 – 2021). Fortunately, there was a small bouncing in 2022 despite of Indonesia was in the attempt to recover from the pandemic. Note, the trends was generated using the final dataset but without those from the jobstreet.com's (P3). P3 was removed because it only contained job postings in the last 30 days (December) and not a full year.

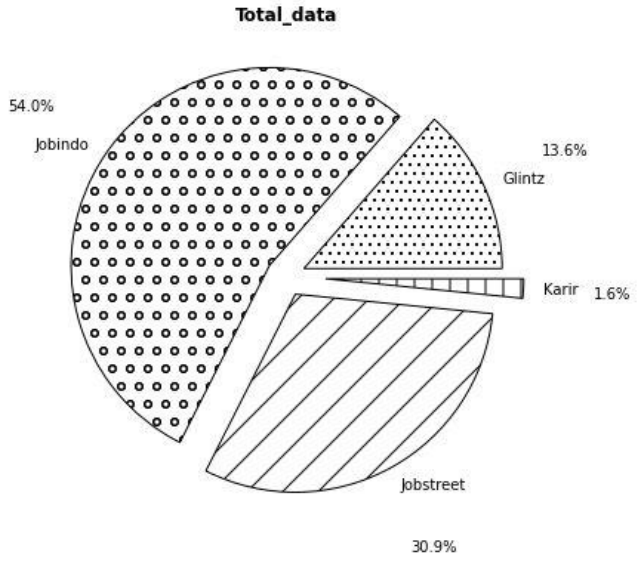

Fig. 4 Platforms record contribution to the dataset

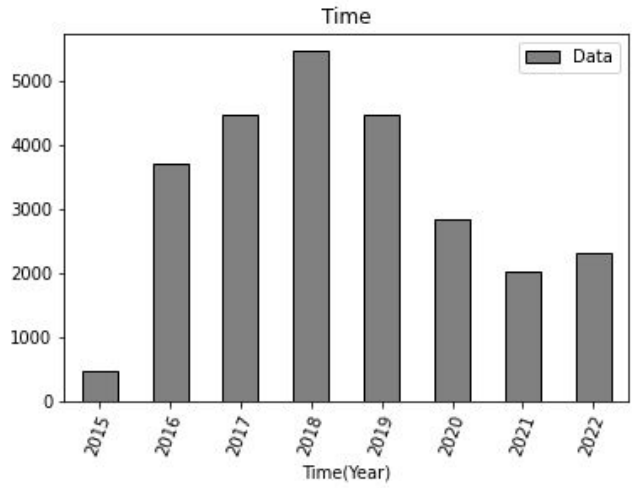

Fig. 5 Job posting from 2015-2022 (without P3)

However, when we incorporated the P3 to generate the trends, the same theme appeared (Fig. 6).

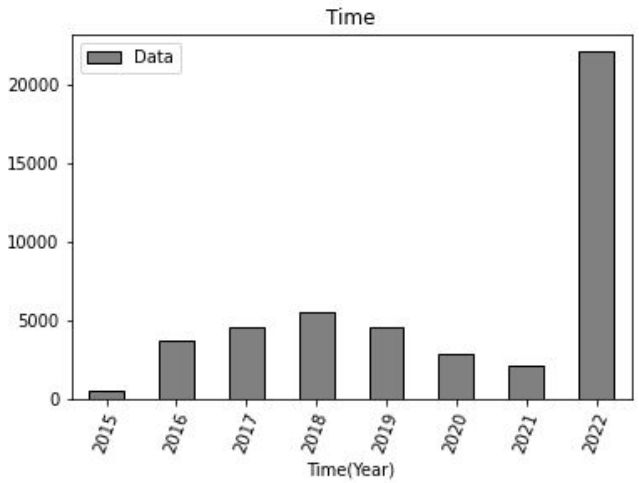

Fig. 6 Job posting from 2015-2022

IV. DISCUSSION

This research collected 63,870 job vacancy records from various types of occupations and countries using web scraping and human annotation techniques. Web scraping techniques take data from job posting websites and convert it into a format that can be accessed and analyzed. However, because the data collected automatically may not be complete or accurate, human annotations are performed to correct and add information to the data that has been collected.

Ref. [1], the research focuses more on data on IT job vacancies in Iran. They collected 7,512 Online IT job vacancies in Iran covering attributes like position, company, location, job description, qualifications, type of contract, salary and publication date. The collected dataset is in JSON format and available on GitHub for easy access by researchers. The data collection process in [1] is carried out by directly accessing job vacancy platforms in Iran and using web scraping techniques. However, in contrast to the first

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

study, the paper [1] uses a crowdsourcing method to perform human annotations on the data that has been collected.

Meanwhile, In [6], the research analyzes the concentration of the labor market in the United States, using data on job vacancies dared. Although it did not mention the specific amount of data or the data format used, this research covered job categories, locations, qualifications, and companies posting job vacancies. However, it is not explained whether the dataset used is publicly accessible or not. The data collection technique used in this study was also not describe in detail.

V. CONCLUSION

Job vacancy opening is an indication of the economic condition of a country. When there is more job openings then the economy of the country is getting better. In this paper we have shared our job vacancy dataset, gathered from four major platforms in Indonesia. We also have described in detail the steps in selecting the platforms, gathering the data, and pre-processing them.

The final dataset constitutes of 63,870 records with 20 attributes from 2015 – 2022. This dataset can be used to assist future research in many areas. Based on the dataset we may create a job recommendation system, identifying trends, making future predictions, field clustering, and many more.

Related to our initial question on how far the global recession would affect Indonesia seen from the job postings. From our result, we can say that the global recession does not clearly affect the job postings. The trends (Fig. 5 and Fig. 6) indicates a mild uplift in job postings, meaning the economics was improving. However, we believe that adding the 2023 dataset, at least Q1, will generate a more in depth insights.

Acknowledgment

This publication is fully supported by Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Institut Teknologi Del with contract number 009.16/ITDel/LPPM/ Penelitian/III/2023.

REFERENCES

[1] F. Noorbehbahani, N. Akbarpour, and M. R. Saeidi, “IranITJobs2021: a Dataset for Analyzing Iranian Online IT Job Advertisements Collected Using a New Crowdsourcing-based Dataset Gathering Process,” presented at the 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2022.

[2] F. H. Sulaeman, “Global recession impacts new ventures in Indonesia: McKinsey,” *The Jakarta Post*, New York, United States, Dec. 2022.

[3] S. Vladimir, “Shopee succumbs to Indonesia start-up layoffs wave, blames global economy,” *The Jakarta Post*, Jakarta, Sep. 2022.

[4] W. D. Herlinda, “Indonesia’s ‘tech winter’ continues as Ajaib announces layoffs,” *The Jakarta Post*, Jakarta, Nov. 2022.

[5] Anika, “Applying Data Mining For Job Recommendations By Exploring Job Preferences,” Thapar University, Palatia, India, 2014.

[6] J. Azar, I. Marinescu, M. Steinbaum, and B. Taska, “Concentration in US labor markets: Evidence from online vacancy data,” *Elsevier*, 2020.

[7] S. Garcia, J. Luengo, and F. Herrera, *Data Prepocessing in Data Mining*. Granada: Springer, 2014.

[8] J. Han and M. Kamber, *Data mining: Concept and Techniques*. San Francisco: Elsevier Inc., 2006.