International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI),
Vol.11, No.3, August 2022

JOB MATCHING USING ARTIFICIAL INTELLIGENCE

Sukit Kitichalermkiat[1], Vishakha Singh[2], Suntaree Chaowiang[2], Sumet Tangprasert[2], Pornthep Chiraprawattrakun[2] and Nopbhorn Leeprechanon[1]

[1]Department of Electrical and Computer Engineering, Thammasat University, Thailand
[2]Depth First Co. Ltd., Bangkok, Thailand

ABSTRACT

The job market has expanded exponentially in the past few years. With many recruiters and candidates, it is not an easy task to match a perfect candidate with a perfect job. The recruiter targets candidates with the required skill sets mentioned in the job descriptions, while candidates target their dream jobs. The search frictions and skills mismatch are persistent problems. In this paper, we build a model that would match companies with candidates with the right skills and workers with the right company. We have further developed an algorithm to investigate people's hiring history for better results.

KEYWORDS

Job matching, Recruiter, TSIC, TSCO, Machine Learning

1. INTRODUCTION

Classifications are not effortless, but the complexity of document/text classification increases two-fold. Generally, the classification approach of text/image/video consists of defining the training data into different classes or categories for training and then using the trained model for classifying the new unseen data into their specified category/class. The data has also increased in the past decade, making it even harder to classify by a human. The job sector has seen tremendous growth in the past few years, and as we see currently with covid 19, the job postings and job searches are only done through the internet.

When it comes to job title, it can sometimes be very vague. For example, a computer engineer does the same job as a web developer. Python programmer does the same job as a python developer, which is quite similar to a python software engineer, but one may be faster and stronger than the other. Therefore, when a recruiter uploads a job posting, it would want to see all the available candidates that possess the skills that are needed for the role. The same applies to the candidates as well. Also, many job titles are written differently but hold the same requirements. It may make the perfect candidate miss that because of the misunderstanding of the job title. Therefore, it is of utmost importance that we have a system that can classify these titles based on their requirements which can benefit the company and the candidate in the long run. It would also be a fairer platform where the candidates will be shortlisted purely based on their skill set.

In Thailand, jobs/occupations are divided into various categories based on the job's skills and requirements. They are regarded through the Thailand Standard Classification of Occupations (TSCO) and the Thailand Standard Industrial Classification (TSIC). In this paper, we propose a job matching classification system that would use the job description alongside TSCO and TSIC to classify the incoming jobs into their specific category. To do the classification job, we used the

DOI : 10.5121/ijscai.2021.11301

Stochastic Gradient Descent, a supervised learning algorithm in our research. Furthermore, we have developed our algorithm, Hiring History, which will review a person's previous hiring details.

2. LITERATURE

2.1. Review

Rodrigues and Chavez [1] in their paper described job matching to be a way to control exactly how a job register and a job candidate are correctly paired together. Their paper extracted the attributes from the resumes and then used WEKA for data mining, they then adopted clustering algorithm to match the profile of the job seekers against the job requirements given by the respective companies.

Almalis Et al. [2], proposed a content-based recommendation algorithm called FoDRA (Four Dimensions Recommendation Algorithm) where they use a structured form of the job and candidate's profile that were produced from the job description and CVs to assess the appropriateness of a job seeker for a particular job position.

Harris [3], in his paper evaluated three approaches to find best candidates to match a set of job skills. He used crowdworkers in a gamified environment, information retrieval-based search methods and a text-mining approach that used feature and elements from the IR-based search engine. He found that the crowdsourcing environment provided the best results for the technical job postings and the crowd and text-mining both performed equally well for the non-technical job postings.

Chalidabhongse, Jirapokakul and Chutivisarn [4] proposed a decision support system called Job Application Support System to facilitate the recruitment process where they focused on the part where the applicants have to fill out application forms and the screening process.

Mishra, Rodrigues and Portillo [5] in their paper "An AI Based Talent Acquisition and Benchmarking for Job" proposed a methodology to solve problem of selecting best CV from a pool of CVs by matching the skill graph generated from CV and Job Post. Their approach is to understand the business aspect to explain why these kinds of problem generate and how one can solve it using natural language processing and machine learning techniques.

Koh and Chew [6] in their paper proposed an intelligent job matching with self-learning recommendation engine for the self-operation of resume matching/ranking. Their parameters include domain of job, job title, position, knowledge, experience, location, salary and other. Their engine is going to extract the data from ontology to ensure the data stability.

Lee, Kim and Na [7], in "Artificial Intelligence based Career Matching" developed a method for career matching amidst university students and companies by the name of Artificial Intelligence based Design platform (AID). They analysed the results from the model with statistical methods like least squares, Pearson correlation, Manhattan distance. In their experimentation they found that their model/methods gave them zero miss-matching between student's skills and company's need on the other hand statistical method gave 30% miss-matching.

2.2. Natural Language Processing

We as a human species mainly communicate with each other via text or speech. We see texts wherever we go from road signs, news outlets, emails, messages, to menus and instructions, that is naturally how we communicate around the world. Natural language processing or NLP can be defined as a branch of Artificial intelligence that gives computers the ability to understand the text and speech like the humans do. In other words, it is used to describe the way of a software's automatic manipulation of these natural languages (speech or text) [8] [9].

In the recent times. NLP has grown popular and can be seen everywhere, from the translation software, GPS systems to chat bots. It's still emerging area and is one of the most complicated areas to tackle in the AI world, as the data has increased tremendously and the fact that human language is filled with ambiguities. But today's technology and high computational machines analyses more data (language based) that humans in an unbiased way and they work tirelessly [10].

NLP can further be divided into different tasks ranging from speech recognition, word sense disambiguation named, entity recognition, sentiment analysis to natural language generation. In this paper, we are attempting to solve the challenges faced by the job sector by applying AI to it.

2.3. Stochastic Gradient Descent (SGD)

Gradient Descent is one of the most popular optimization techniques in Machine learning and Deep Learning and can be used with most of the learning algorithms. The gradient or slope is the rate of change of a function and is typically used for functions that has several inputs but a single output.

For our work, we have chosen Stochastic Gradient Descent (SGD), as it is one of the most efficient approaches to fit linear classifiers and regressors under covex loss functions such as linear Support Vector Machines and Logistic Regression. The word 'stochastic' refers to a system that is related with random probability, that is, this algorithm selects samples at random and not the whole data for training for each iteration [11]. It has gotten more famous in the recent times due to the large-scale learning and is often used for machine learning problems that are faced in text classification and natural language processing (NLP). SGD is easy to implement and is efficient but can be sensitive to feature scaling.

2.4. TSCO

TSCO stands for Thailand Standard Classification of Occupations (TSCO-2001), it is inspired from ISCO (International Standard Classification of Occupations) where according to the skill level occupations (Jobs) are divided into various categories ranging from low to high and armed forces [12]. They are further divided into Major groups where there are total of 9 categories, from Managers, Technicians and Associate professionals to Elementary Occupations and Armed Forces Occupations. Influenced by ISCO, Thailand in 2001 made their own Standard Classification and termed it TSCO, which is what we have used in our paper.

2.5. TSIC

Similar to TSCO, there is another set of industrial classification known as TSIC (Thailand Standard Industrial Classification). Inspired from ISIC (International Standard Industrial Classification of all Economic Activities) which is a United Nations system for classifying

different economic data. Various kinds of economic activity in the fields of production, employment, gross domestic product and other statistical areas are classified here [13]. TSIC consists of 21 sections, 88 divisions, 243 groups, 440 class and 1089 industries in total. Ranging from agriculture, forestry, construction to manufacturing, mining etc.

In our paper we have made use of TSCO and TSIC for the classification of different occupations/jobs.

3. METHODOLOGY

The goal of our work is to get a classifier that receives a job position which would be in a form of a sentence, written in natural language (English and Thai) and the return the TSCO code for that job position.

Our work is carried out using the Python language and the implementation of the trained classifiers is done in Scikit Learn library, and classification is done using SGD classifiers.

3.1. Data Processing

The classifier we have used in our work belongs to supervised learning, and they are trained on the data that we have collected from Department of Employment/Labour and are manually classified. The input data in our work is job descriptions and job titles with their respective TSCO. The aim is for the model to learn the TSCO for different jobs and when model is fed with unseen data, they map the TSCO to their categories.

The input data looks like the following data shown in the table 1, which is provided to the model in a tab separated file:

Table 1. Input dataset for the model

Position	**Classification**
ผู้ปฏิบัติงานอาชีพในกลุ่มนี้รวมถึงนักบริหารของหน่วยงานรัฐบาลซึ่งมิได้จัดประเภทไว้ในที่อื่นกลุ่มอาชีพในหมู่นี้มีดังนี้	1120
ควบคุมการรับ– จ่ายเงินการบัญชีและงบดุลของพรรคตามที่กฎหมายหรือตามที่คณะกรรมการบริหารพรรคกำหนด	1141
ปฏิบัติงานหลักมูลฐานเช่นเดียวกันกับกำนันแต่ปกครองบรรดาราษฎรที่อยู่ในเขตหมู่บ้านและดำเนินการในเรื่องต่างๆตามที่กำนันมอบหมาย	1130
ผู้ปฏิบัติงานอาชีพในกลุ่มนี้รวมถึงผู้จัดการด้านภัตตาคารและโรงแรมซึ่งมิได้จัดประเภทไว้ในที่อื่นอาชีพในหน่วยนี้มีดังนี้	1225
ผู้ปฏิบัติงานอาชีพในกลุ่มนี้รวมถึงผู้จัดการด้านการขนส่งการสื่อสารและคมนาคมซึ่งมิได้จัดประเภทไว้ในที่อื่น	1226
หรือสถานประกอบการอื่นๆ	1233
ใจว่าได้รับมอบตามกำหนดเวลา	1235
ลอบตัดต้นไม้	1311
ผู้ปฏิบัติงานอาชีพในกลุ่มนี้รวมถึงผู้ประกอบวิชาชีพด้านคอมพิวเตอร์ซึ่งมิได้จัดประเภทไว้ในที่อื่นกลุ่มอาชีพในหมู่นี้มีดังนี้	2139

The first column refers to the <job description> and the second column defines the respective <job position> (TSCO code). Since the classifier is a supervised learning algorithm, we are training them using manually classified data.

We then generated dataframe for the given input file that has the job description and the TSCO code. After which splitting is done for training (train(x)) and testing (test(y)). Normalization of the dataframe is done according to our defined criteria, the results from the above process gives us the X_train, X_test and Y_train, Y_test sets. So, this step splits our input data into training (80%) and testing set (20%).

3.2. Training and Tuning our Model

SGD classifiers consists of parameters and hyper-parameters. Parameters are the values that corresponds to the mathematical model, that are adjusted after the training while, hyper-parameters are values that are relate to the way of training and are adjusted using the selected part of the training set. Learning rate, momentum, decay and nesterov are the hyperparameters that can be optimized in SGD.

To train and adjust the parameters is feasible to use a fit function, in our work we have made use of tools provided by Scikit learn known as Pipeline and GridSearchCV with the aim to make extensive search to achieve the hyperparameters needed to optimize the results [14]. Below table 1 and 2 shows the configuration of SGD classifier and GridSearchCV used in our paper.

Table 2. Parameters of SGD Classifier

Parameters	Definition	Value
Leaning rate	To control the way an algorithm learns and to help the performance through tuning	0.01
loss	To evaluating how well the algorithm is modelling the given data	'hinge'
fit_intercept	Whether the intercept should be estimated or not.	True
max_iter	The max number of passes over the training data.	1000
l1_ratio	It is the Elastic Net Mixing parameter. with 0 <= l1_ratio <= 1. l1_ratio=0 corresponds to L2 penalty, l1_ratio=1 to L1.	0.15
Penalty	It is the total of absolute values of weights.	'l2'
Shuffle	To determine if the training data should be shuffled or not.	True
power_t	To inverse scale the learning rate.	0.5
validation_fraction	The amount of data set aside for validation.	0.1
class_weight	Weights that are associated with the given classes.	'balanced'
Tol	The stopping criterion.	1e-3
epsilon	It is used when there is a selection of specific action based on the Q values we already have.	0.1

Table 3. Parameters of GridSearchCV

Parameters	Definition	Value
n_jobs	To determine the number of jobs that run in parallel	2
cv	To determine the cross-validation splitting strategy	5
iid	To assume if the data is identically distributed across the folds or not.	false

3.3. Our Algorithm (Hiring History Algorithm)

We have developed a model that after the job matching model is done will look into a person's hiring history for better prospect. This algorithm improves the matching score and can be used for further personalization and the idea is adopted from Discounted Cumulative Gain. Our algorithm can be expressed as below:

Rating (R)

$$R_i = \frac{\left(\frac{t_i}{\sum_{n=1}^{N} t_n}\right)}{\log\left(1+\sum_{n=1}^{i} t_n\right)} \quad \text{Eq. 1}$$

R = Rating
i = the i^{th} job in the candidate's work history
t = Time (how much time the candidate worked)
N = total number of jobs in the candidate's work history

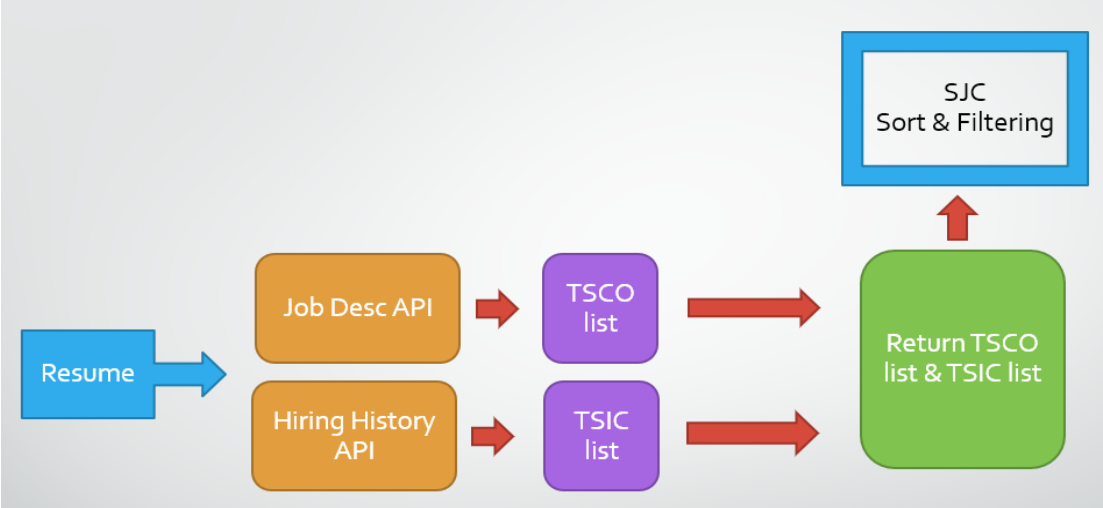

Fig 1. Job matching API to find job

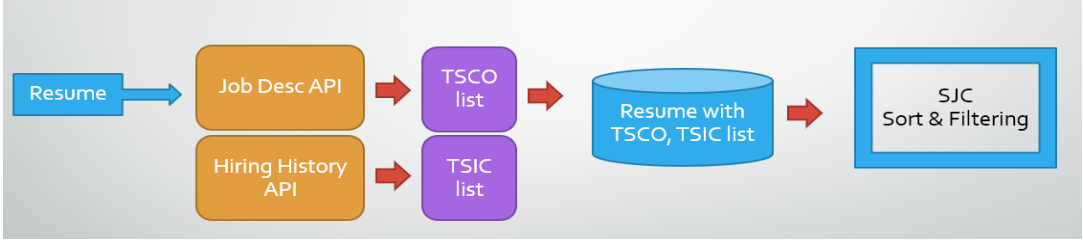

Fig 2. Job matching API to find candidate

Figure 1 and 2 shows the way the information is carried out in our model. The algorithm is tested on the data gathered from the Department of Employment. The model is already in use throughout Thailand, the steps and working of the algorithm are as follows:

1. Firstly, the algorithm will look into the candidate's current and previous job profile, i.e., where they are currently working, where in the past they have worked and what kind of job positions were they holding at that time. Fig 3 shows layout of it.

Industry(i)	TSIC	time (t)	Rating (R)
1-current job	1	1 year	0.097859
2-previous job	2	3 months	0.023263
3-previous job	2	6 months	0.044128
4-previous job	1	2 year	0.160015
5-previous job	3	3 months	0.019555

Fig 3. Job history of a candidate A

2. After that, we will calculate the rating for each job to evaluate which job industry is better suited for that particular candidate.
3. Higher rating will be given to the current job position and the score will keep on getting lower with the past jobs. Fig 4 predicts the rating for respective job position.

TSIC	Rating (R)	%
1	0.257873	75
2	0.067391	19
3	0.019555	6

Fig 4. Rating for the job best suitable for candidate A.

4. Finally, a smart job center (SJC) will filter and sort the jobs in its database according to the suggested TSCO and TSIC.
5. With this, our algorithm is able to determine which sector of occupation is best for the given candidate.

4. Results and Confusion Matrix

As we are aware, the NLP is of the most complex of AI problems. In this paper, we attempted to tackle the job matching problem with Stochastic gradient descent and we received the overall accuracy of 99.89%. Our result also gives macro average and weighted average, as macro average gives each prediction equivalent weight when it is calculating but it may sometime occur that our data might be imbalanced and we would like to give importance and value to some prediction more, there we need weighted average. So, our results compromises of all the given accuracies.

accuracy	0.998964	0.998964	0.998964	0.998964
macro avg	0.999332	0.999246	0.999264	7721
weighted avg	0.999015	0.998964	0.998962	7721

Fig 5. Accuracies of the job matching model

Table 4. Testing result of the model

Position	Classification
เจ้าหน้าที่ดูแลเอกสารด้านกฎหมาย	M72
นักศึกษาฝึกงาน E-Commerce	M742
เจ้าหน้าที่ Online Marketing	S9610
Supervisor Laboratory ประจำโรงพยาบาลศรีระยอง	M72
หัวหน้าแผนก QA	G471
Creative Content & Copy Writer	M742
เจ้าหน้าที่แมสเซนเจอร์	M72

The prediction results from our model on an unseen data looks like above data represented in the table 4. The left side represents the job title and the right side depicts the TSCO. We provided the model with only job titles and got their respective TSCO in return.

Figure 6 shows the confusion matrix where various job positions were classifies into their respective categories. A part of our confusion matrix for job matching model can be seen in fig 6 below. The horizontal line is our True Class whereas the vertical line is our Predicted Class, and it is done in order to visualize our classification algorithm's performance. As we can see from figure below, our algorithm is able to predict the true class with good accuracy,

	A	B	C	D	E	F	G	H	I	J
1		**0**	**1000**	**1100**	**1110**	**1120**	**1130**	**1140**	**1141**	**1142**
2	**0**	3	0	0	0	0	0	0	0	0
3	**1000**	0	3	0	0	0	0	0	0	0
4	**1100**	0	0	3	0	0	0	0	0	0
5	**1110**	0	0	0	25	0	0	0	0	0
6	**1120**	0	0	0	0	19	0	0	0	0
7	**1130**	0	0	0	0	0	28	0	0	0
8	**1140**	0	0	0	0	0	0	3	0	0
9	**1141**	0	0	0	0	0	0	0	25	0
10	**1142**	0	0	0	0	0	0	0	0	6
11	**1143**	0	0	0	0	0	0	0	0	0
12	**1144**	0	0	0	0	0	0	0	0	0
13	**1150**	0	0	0	0	0	0	0	0	0
14	**1200**	0	0	0	0	0	0	0	0	0
15	**1210**	0	0	0	0	0	0	0	0	0
16	**1220**	0	0	0	0	0	0	0	0	0
17	**1221**	0	0	0	0	0	0	0	0	0

Fig 6. Confusion matrix

4.1. Hiring History Algorithm

The algorithm, test on the data from Department of Employment show how well it will work for the future usage. Hiring History algorithm gives the accuracy of 73.81% and precision of 77.02%. The TSCO code is the result that we receive from our above job matching model, and the TSIC code is the result from our hiring history algorithm. In Figure 7, it shows that when a person looking for a job in Human Resources field, the algorithm will return all jobs with TSCO 1232, which is Personnel and Industrial Relations field, and TSIC 46599 (Machine & Computer trading), 47412 (Video Games and Software) and 47112 (Discount Store and Hyper market) respectively according to the person's working history.

```
{
    "text": "ฝ่ายบุคคล",
    "id": "3101800943664",
    "result": [
        {
            "tsco": "1232",
            "tsic": "46599",
            "score": "0.96111733"
        },
        {
            "tsco": "1232",
            "tsic": "47412",
            "score": "0.92142982"
        },
        {
            "tsco": "1232",
            "tsic": "47112",
            "score": "0.40435364"
        },
        {
```


Fig 7. Testing result of Hiring History algo

Similarly, confusion matrix for the above test data is given in fig 8 where, the predicted class and true class matrix is defined. The red diagonal numbers refer to the number of correct predictions whereas the yellow highlighted number explains which of true class were predicted wrong.

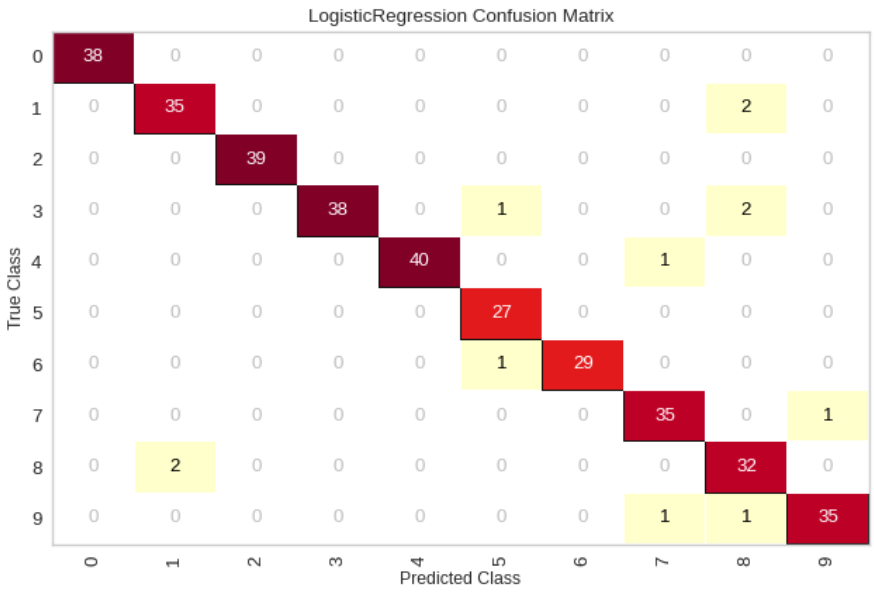

Fig 8. Confusion matrix of Hiring History

5. CONCLUSIONS

In this paper, we have used a job matching model based on stochastic gradient descent to solve the problem occurring in the job sector. We have further developed an algorithm to further enhance our model. The algorithm takes a review at the people's previous hiring setting, and provides the best suitable candidate for a particular job profile. The model is currently being used internally by the Department of Employment, Thailand.

6. LIMITATIONS AND FUTURE WORK

First limitations we faced was due to the nature of dataset necessary for this experiment. For training and testing we needed personal information of people in regard to their work which makes it difficult to get vast dataset. Along with that we can never be sure of the information received from candidates, if the satisfaction level stated by them is true or otherwise. Another limitation is that sometimes our model will suggest a former job that might not be available anymore.

Overall limitation of our work would only be the dataset. Our model can be used in other countries if the department that is using this have a right to access their people's personal data.
Future work of our model will depend on the suggestions we will receive by the Department of Employment, Thailand. We would like to add more attributes that is responsible in a hiring process. We will try to gather the data from different companies and will look into their hiring process, their requirements and other internal managements and will add those to our existing model.

ACKNOWLEDGEMENTS

Financial support granted by Depth First Co, Ltd. (Thailand) is greatly appreciated and acknowledged.

REFERENCES

[1] Rodriguez, L. G., & Chavez, E. P. (2019). Feature selection for job matching application using profile matching model. *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. https://doi.org/10.1109/ccoms.2019.8821682

[2] Almalis, N. D., Tsihrintzis, G. A., Karagiannis, N., & Strati, A. D. (2015). FoDRA — A new content-based job recommendation algorithm for job seeking and recruiting. *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*. https://doi.org/10.1109/iisa.2015.7388018

[3] Harris, C. G. (2017). Finding the best job applicants for a job posting: A comparison of human resources search strategies. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw.2017.31

[4] Chalidabhongse, J., Jirapokakul, N., & Chutivisarn, R. (2006). Facilitating job recruitment process through job application support system. *2006 IEEE International Conference on Management of Innovation and Technology*. https://doi.org/10.1109/icmit.2006.262244

[5] Mishra, R., Rodriguez, R., & Potillo, V. (2020). An AI Based Talent Acquisition and Benchmarking for Job.

[6] Koh, M. F., & Chew, Y. C. (2015). Intelligent job matching with self-learning recommendation engine. *Procedia Manufacturing*, *3*, 1959-1965. https://doi.org/10.1016/j.promfg.2015.07.241

[7] Lee, D., Kim, M., & Na, I. (2018). Artificial Intelligence based career matching. *Journal of Intelligent & Fuzzy Systems*, 35*(6), 6061-6070*. https://doi.org/10.3233/jifs-169846

[8] *What is natural language processing?* (2019, August 7). Machine Learning Mastery. https://machinelearningmastery.com/natural-language-processing/

[9] *What is natural language processing?* (2020, July 2). IBM - United States. https://www.ibm.com/cloud/learn/natural-language-processing

[10] *What is natural language processing?* (n.d.). https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

[11] *ML | Stochastic gradient descent (SGD)*. (2021, September 13). GeeksforGeeks. https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/

[12] *International standard classification of occupations (ISCO)*. (n.d.). ILOSTAT. https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

[13] *Thailand Standard Industrial Classification: TSIC-2009*. Employment Promotion Division Department of Employment Ministry of Labour MAY 22, 2014.

[14] *S*cikit-learn. Retrieved December 7, 2021, from https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model