



IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

www.ijitce.com



Email : ijitce.editor@gmail.com or editor@ijitce.com

A TWO-STAGE FRAMEWORK FOR ACCURATE JOB TITLE CLASSIFICATION IN ONLINE ADVERTISEMENTS

¹Pothula Lakshmi Narasimha, MCA Student, Department of MCA

²P. Rohini Bai, M.Tech, Assistant Professor, Department of MCA

¹²Dr KV Subba Reddy Institute of Technology, Dupadu, Kurnool

ABSTRACT

Large databases may be mined for knowledge using data science approaches. Recently, there has been a lot of interest in categorising online job advertisements (ads) in order to analyse the labour market. To determine the occupation from a job advertising, many multi-label classification techniques (such as self-supervised learning and clustering) have been developed and have shown satisfactory results. Nevertheless, these methods rely on specialised databases like the Occupational Information Network (O*NET) that are more suited to the US labour market and need for labelled datasets with hundreds of thousands of samples. To solve the issue of tiny datasets, we introduce a two-stage job title identification mechanism in this study. First, we categorise the job advertising by sector (e.g., Agriculture, Information Technology) using Bidirectional Encoder Representations from Transformers (BERT). The closest matching job title is then identified from the list of jobs within the anticipated sector using unsupervised machine learning methods and a few similarity metrics. In order to solve the problems of processing and categorising employment advertisements, we also suggest a unique document embedding technique. According to our experimental

findings, the suggested two-stage method increases the accuracy of job title detection by 14%, reaching over 85% in some industries. Furthermore, we discovered that, in comparison to methods based on the Bag of Words model, integrating document embedding-based techniques such as weighting schemes and noise reduction increases the classification accuracy by 23.5%. Additional assessments confirm that the suggested methodology either surpasses or performs on par with the state-of-the-art techniques. Finding new and in-demand jobs in Morocco has been made easier by applying the suggested technique to data from the Moroccan labour market.

1. INTRODUCTION

Due to the digitisation of processes and the growth of social media, the Internet is now widely used in many sectors. This has led to a vast volume of data that must be processed and analysed quickly and efficiently in order to extract useful insights that may aid in decision-making [1]. In this regard, data science methods can be effective instruments for obtaining information from sizable datasets, aiding in the classification of various data types (such as text, images, and video) [2], and resolving numerous other issues that are addressed in a conventional way, which frequently requires a lot of time and resources.

In a similar vein, internet job portals and websites replaced traditional job market methods. This is due to the fact that in order to reach a wider audience and target more job searchers, recruiters and companies post different job openings on numerous platforms. Many stakeholders can gain from this change as it offers a chance to comprehend the demands of the labour market from the massive volume of data that is exchanged every day [3]. Specifically, determining the skills and professions needed can assist policymakers and labour market analysts in promoting employment, as well as in helping students and job seekers locate appropriate positions and the training they need to make a smooth transition to the workforce [4].

It's not an easy undertaking to categorise internet job adverts. In fact, a job ad's content is presented in plain language in a semi-structured or unstructured fashion, and the terminology that employers use in the text frequently differs greatly from the occupational classifiers and databases created by human resources specialists. Furthermore, job advertisements frequently contain material that is too general and unrelated to the role. This complicates the process of connecting the job posting to the appropriate occupation. For example, the title of a job posting could contain details like the city in which the position is situated or a wage range. Information about the business and other duties that aren't directly related to the desired job may also be included in the description. To overcome these obstacles, sophisticated word and

document representation approaches as well as innovative feature extraction techniques must be used.

The majority of suggested approaches see occupation normalisation as a classification or grouping issue. Various text classifiers, including support vector machines (SVM) [1], naïve bayes [5], k-nearestneighbor (KNN) [1], [5], artificial neural networks (ANNs) [6], and Bidirectional Encoder Representations from Transformers (BERT) [7], have been proposed for this task, spanning from traditional machine learning (ML) models to deep learning models. The authors in [8] utilised just the title and discovered that 30% of the job offer titles lacked sufficient information to identify the occupation, but other research used both the title and the description to do classification. In a similar vein, the authors in [9] discovered that one job description may correlate to many occupations after focussing just on the job description language. To the best of our knowledge, no prior research has looked at the role that the title and description play in normalising job advertisements. Using internal taxonomy or an occupational classifier to categorise job advertisements has typically produced satisfactory results [6], [10]. Nevertheless, these approaches need hundreds of thousands of samples in human-labeled datasets, which takes a lot of time and resources. Additionally, because the entire training process needs to be performed, it is exceedingly challenging to update the occupation description or add a newly generated occupation to the occupational classifier. Furthermore, it is

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

difficult to extend previous work to job advertising written in other languages because the majority of it primarily concentrates on English-language job ads and makes use of certain occupational classifiers like the Occupational Information Network (O*NET). It makes translating their methods into other languages very challenging.

However, use unsupervised methods to determine the occupation—like clustering [11] and field similarity [12]—avoids training the model with labelled data, which isn't always accessible. This is especially important because there are a lot of jobs to consider. In most earlier publications, word embedding was generated using basic approaches like Bag of Words (BOW) [1], [12], or Term Frequency Inverse Document Frequency (TFIDF) [11], and the document embedding was calculated using averaging methods. When dealing with job advertisements published by various employers who use distinct lexicons, these methods are seen to be inadequate in capturing the semantic links between the terms. Because state-of-the-art methods don't always work effectively, word embedding strategies and feature extraction procedures [13] must be constantly watched to get the best outcomes [11].

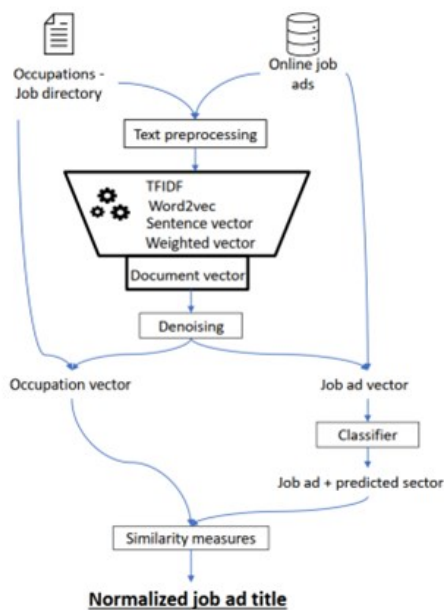
In order to get beyond the aforementioned drawbacks, we provide in this study a job title recognition approach based on self-supervised and unsupervised machine learning algorithms with high accuracy and little labelling that can be applied to data from various nations. Two

parts make up the suggested methodology: first, job advertising are categorised by sector, and then they are matched with vocations that fall within the anticipated sector. In order to classify job ads into their respective sectors (such as Information Technology (IT), Agriculture), a number of text classifiers, including SVM, Naïve Bayes, Logistic Regression, and BERT, are used. This allows us to concentrate on the occupations of the predicted sector rather than using all of the occupations from the occupational classifier. In order to suggest a personalised document embedding approach for the job title identification stage, we examine several methods for vector representation of texts and employ a number of combinations and factors. Additionally, we explore a number of feature selection techniques to extract key terms from the description and assess how much the title and description contribute to better outcomes. In order to select the closest representation, we lastly compute the similarity between the job ad representation and the occupation representations that correspond to the anticipated sector. In order to accomplish this, we gather around two hundred thousand job advertisements from employment sites as well as the French occupational classifier "Pole Emploi." Our approach yields an overall accuracy of 76.5% and more than 85% for some sectors when used to identify the occupation title on a random sample of job advertising, which is regarded as high accuracy in comparison to previous studies. Additionally, a group of domain experts manually labelled a sample of our dataset to confirm the efficacy of our methodology. Lastly, in order to obtain a

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

general picture of the Moroccan labour market, particularly in the IT industry, we applied our technique to a dataset of 248,059 French-language job advertisements. This study enables us to clarify important industries and professions in the Moroccan labour market, where a prior study on the country's offshore industry found a significant need for IT profiles and telemarketers [14]. By using this approach, we can find new professions that can assist decision-makers, such as academic institutions, in modifying their programs and curriculum and in assisting students and job seekers in orienting themselves to a professional path that will lead to employment [4].

2. SYSTEM ARCHITECTURE



3. EXISTING SYSTEM

Before determining the necessary skills based on job responsibilities, several research have tried to standardise job ad names as a first step in constructing job

advertising. A collection of occupations with comparable duties and skill requirements is called an occupation. It is crucial to remember that jobs and job titles are not the same as vocations. A job is associated with a particular work environment and is performed by a single individual, whereas occupations aggregate jobs according to shared traits. By deriving skills from structured skill bases that include complete occupation descriptions like the International Standard Classification of Occupations (ISCO) or O*NET, determining the necessary occupations in the labour market can be viewed as a top-down method of learning the necessary skills.

There are two methods for figuring out job titles from job postings. The first method classifies job titles using supervised models, whereas the second method finds the closest job title using unsupervised models. This section examines earlier research on job ad categorisation techniques. Many research used SVM and KNN to classify job advertising according to the standard referential, framing the goal of job title recognition as a text classification task. Particularly in [1] and [18], CareerBuilder.com employed a multi-stage classifier to handle a large number of classes, which is nearly identical to the application domain (online recruitment) utilised by LinkedIn's job title classification system [15], which relies heavily on crowdsourced labelling of training samples and a heavily manual phrase-based classification system reliant on short-text. Additionally, in [16], they used string similarity to train the siamese network to categorise job names by

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

feeding it comparable job titles. They classified the job titles for this work using an internal taxonomy rather than O*NET and ISCO bases. Additionally, text classifiers were utilised in [6, 7], [8, and 10], ranging from conventional machine learning models to deep learning models based on ISCO occupation classifiers or on customised lists of occupations, respectively. While text classifiers in [6] and [10] performed well in extracting the abilities required for certain jobs, those in [8] just utilised the job ad's title and omitted the description, which resulted in a less interesting accuracy. In conclusion, the authors found that around 30% of job offer titles lack sufficient information to identify the occupation.

In a related research, the authors classified job titles according to the query description into 30 different classes, which corresponded to the top 30 occupations, using a Kaggle dataset [5]. They employed a number of methods, including Random Forest, Bernoulli's Naïve Bayes, Multinomial Naïve Bayes, and Linear SVM. They discovered that the accuracy of job title classification is improved by expanding the training set, and that Linear SVM produces the best results. The authors of [9] conclude by suggesting a multi-label classification method for identifying pertinent job titles from job description texts, taking into account the possibility that one job description may relate to many occupations. They use several pre-trained language models to apply the technique described in [7] to the problem of job title prediction. They discovered that BERT with a multilingual pre-trained model produced

the best results on their dataset and that additional information, such as the job name, job level, and work criteria, is necessary for the prediction because the description alone is insufficient.

Disadvantages

- The primary drawback of text classifiers is the cost of acquiring data for training with thousands of occupational groups, many of which are not all that different from one another.
- Because there aren't many labelled datasets available for the training phase in an existing system, we decided to combine the two methods.

4. PROPOSED SYSTEM

In order to get beyond the aforementioned drawbacks, we provide in this study a job title recognition approach based on self-supervised and unsupervised machine learning algorithms with high accuracy and little labelling that can be applied to data from various nations. Two parts make up the suggested methodology: first, job advertising are categorised by sector, and then they are matched with vocations that fall within the anticipated sector.

In order to classify job ads into their respective sectors (such as Information Technology (IT), Agriculture), a number of text classifiers, including SVM, Naïve Bayes, Logistic Regression, and BERT, are used. This allows us to concentrate on the occupations of the predicted sector rather than using all of the occupations from the occupational classifier. In order to suggest a personalised document embedding approach

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

for the job title identification stage, we examine several methods for vector representation of texts and employ a number of combinations and factors. Additionally, we explore a number of feature selection techniques to extract key terms from the description and assess how much the title and description contribute to better outcomes.

In order to select the closest representation, we lastly compute the similarity between the job ad representation and the occupation representations that correspond to the anticipated sector. In order to accomplish this, we gather around two hundred thousand job advertisements from employment sites as well as the French occupational classifier "Pole Emploi." Our approach yields an overall accuracy of 76.5% and more than 85% for some sectors when used to identify the occupation title on a random sample of job advertising, which is regarded as high accuracy in comparison to previous studies.

Additionally, a group of domain experts manually labelled a sample of our dataset to confirm the efficacy of our methodology. Lastly, in order to obtain a general picture of the Moroccan labour market, particularly in the IT industry, we applied our technique to a dataset of 248,059 French-language job advertisements. This study enables us to clarify important industries and professions in the Moroccan labour market, where a prior study on the country's offshore industry found a significant need for IT profiles and telemarketers [14]. By using this technique, we can find new professions

that can assist decision-makers, such as universities, in modifying their programs and curriculum and in assisting students and job seekers in orienting themselves to a professional path that will lead to employment [4].

Benefits

- To enable replication for additional languages and nations, we provide an approach for occupation identification in the event that labelled data is scarce.
- To solve the occupation identification challenge, we compare several document representation methodologies and determine how much the job ad's title and description contribute to the matching process.
- In order to alleviate the constraints in this sector, we gather information on the occupation-specific requirements of the Moroccan IT labour market and create a dataset of French-language Moroccan job advertisements.

5. IMPLEMENTATION

Modules description

Service Provider

The Service Provider must use a working user name and password to log in to this module. He can perform many tasks after successfully logging in, including browsing datasets and training and testing datasets. View Results of Trained and Tested Accuracy, View Trained and Tested Accuracy in Bar Chart, View the Job Title Identification Type Ratio, View the Predicted Job Title Identification Type, Get Predicted Data Sets here. View All Remote Users and Job Title Identification Type Ratio Results.

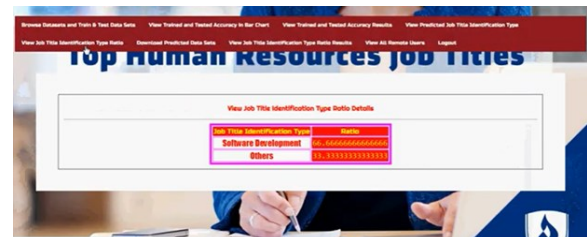
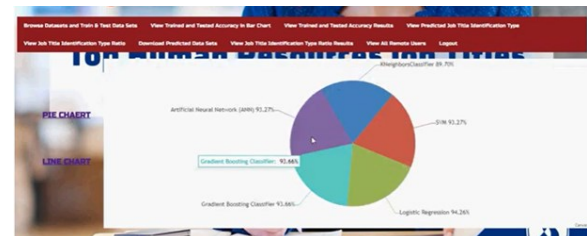
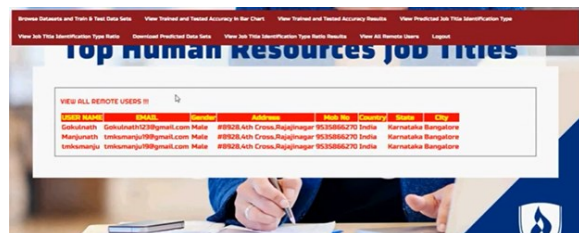
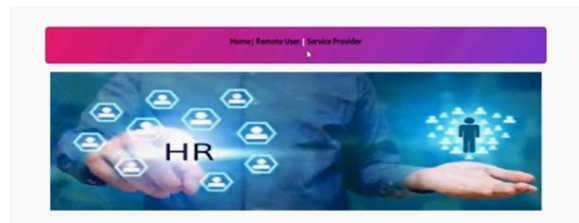
View and Authorize Users

The list of users who have registered may be viewed by the administrator in this module. The administrator can see the user's name, email address, and address in this, and they may also grant the user permissions.

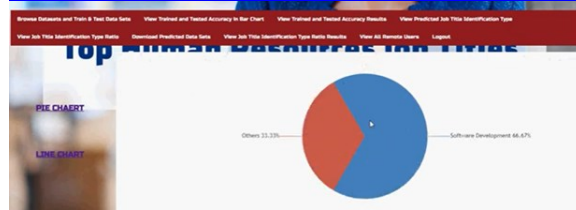
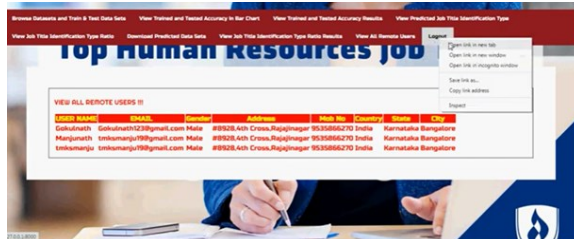
Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user will do tasks such as registering and logging in, predicting the job title and identification type, Examine your profile.

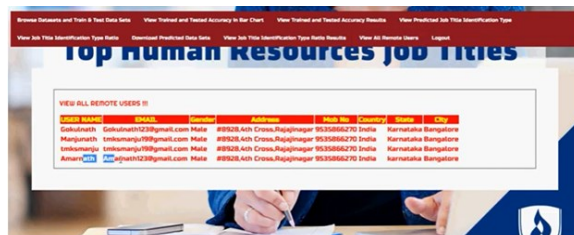
6. SCREENSHOTS



<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>


USER NAME	EMAIL	Gender	Address	Mobile No.	Country	State	City
Gokulnath	Gokulnath123@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Mangunath	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore



USER NAME	EMAIL	Gender	Address	Mobile No.	Country	State	City
Gokulnath	Gokulnath123@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Mangunath	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore



USER NAME	EMAIL	Gender	Address	Mobile No.	Country	State	City
Gokulnath	Gokulnath123@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Mangunath	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore



USER NAME	EMAIL	Gender	Address	Mobile No.	Country	State	City
Gokulnath	Gokulnath123@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Mangunath	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
tsunamang19	tsunamang19@gmail.com	Male	#9928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore

7. CONCLUSION

Our two-stage job tile identification process is based on minimally labelled, semi-supervised, and unsupervised machine learning algorithms. Specifically, we use a typical occupational classifier to choose the most suitable occupation for each job post based on similarity measurements. We

explored a number of word and document representation techniques throughout the tests and after pre-processing the gathered job advertisements, including deep contextualised word representation (BERT), neural language models that rely on distributional semantics (Word2Vec, Fast Text), and TFIDF. All of them underwent a number of weighing techniques to lessen the influence of superfluous words, particularly in the description. Next, in order to determine the extent to which the title and description contribute to the process, we examined a number of balance variables.

Since the similarity measures between the job ad and the occupations would only be used inside the projected sector rather than utilising all of the occupations from the reference, the experiment results showed that categorising the job advertisements by sector increased the accuracy of our methods by 14%. Because the training dataset and job openings had different language, we discovered that W2V produced better results for document representation than BERT. However, we discovered that BERT yields the most accurate answers when the sector is left unspecified. Regarding weighting strategies, the results indicate that the TFIDF weighting strategy significantly improves performance for long text (job ad descriptions, occupation descriptions), while uniform and frequency word weighting is best for short text (job ad titles, occupation titles), which are not sensitive to word weighting. Furthermore, we discovered that, out of all the settings we examined, document embedding utilising only the top N selected words from the description using

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

weighting scores produces the best accurate results since it adds pertinent information to the title. Lastly, trials confirm how well the title and description work together in the matching process. Additionally, they confirm that we shouldn't give them equal weight because the title is more pertinent because it uses more complex terms associated with the position.

We were able to increase our methodology's accuracy by 34% over the baseline thanks to these results. In terms of performance, our outcomes are similar to those of the classification method. In particular, we achieved an overall accuracy of 76.5%, which, depending on the industry, might occasionally surpass 85%. Examples of these industries include the health and hotel and tourist sectors. Moreover, when approaching the task of job title identification as a classification issue, these insights may also be used to increase the classifier's accuracy.

In order to normalise the job advertisements and extract insights from them, this process may be repeated in various languages with minimum intervention using other occupation classifiers. Within the framework of the USAID-supported project "Data science for improved education and employment in Morocco," which aims to analyse job market demands and extract skills from them, the suggested method has been evaluated in a real-world environment [4]. It may also be used when universities are creating training programs based on the demands of the labour market. The findings of research employing this technique to

examine the labour market can also be advantageous to young people and job seekers.

Because recruiters may not adhere to a set structure when creating job advertising, we want to incorporate a phase of job enrichment with skills phrases based on the occupation description in the future to make the job ad and occupation description as comparable as feasible. In order to retain only pertinent terms, we also want to further purify the list of the top N words produced using weighing algorithms. Additionally, we want to use French job-related words to train our own Word2Vec model, which might improve the precision of our approach.

REFERENCES

- [1] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, and T. S. Kang, "Carotene: A job title classification system for the online recruitment domain," in *Proc. IEEE 1st Int. Conf. Big Data Comput. Service Appl.*, Mar. 2015, pp. 286–293.
- [2] M. S. Pera, R. Qumsiyeh, and Y.-K. Ng, "Web-based closed-domain data extraction on online advertisements," *Inf. Syst.*, vol. 38, no. 2, pp. 183–197, Apr. 2013.
- [3] R. Kessler, N. Béchet, M. Roche, J.-M. Torres-Moreno, and M. El-Bèze, "A hybrid approach to managing job offers and candidates," *Inf. Process. Manage.*, vol. 48, no. 6, pp. 1124–1135, Nov. 2012.
- [4] I. Rahhal, K. Carley, K. Ismail, and N. Sbihi, "Education path: Student orientation based on the job market needs," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Mar. 2022, pp. 1365–1373.

<https://doi.org/10.62647/ijitce.2025.v13.i2.pp687-696>

[5] S. Mittal, S. Gupta, K. Sagar, A. Shamma, I. Sahni, and N. Thakur, “A performance comparisons of machine learning classification techniques for job titles using job descriptions,” *SSRN Electron. J.*, 2020. Accessed: Feb. 22, 2023. [Online]. Available:

<https://www.ssrn.com/abstract=3589962>, doi: 10.2139/ssrn.3589962.

[6] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, “Using machine learning for labour market intelligence,” in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science),

[7] T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, “Job prediction: From deep neural network models to applications,” in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Oct. 2020, pp. 1–6.

[8] F. Amato, R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, V. Moscato, F. Persia, and A. Picariello, “Challenge: Processing web texts for classifying job offers,” in *Proc. IEEE 9th Int. Conf. Semantic Comput. (IEEE ICSC)*, Feb. 2015, pp. 460–463.

[9] H. T. Tran, H. H. P. Vo, and S. T. Luu, “Predicting job titles from job descriptions with multi-label text classification,” in *Proc. 8th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Dec. 2021, pp. 513–518.

[10] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, “Classifying online job advertisements through machine learning,” *Future Gener. Comput. Syst.*, vol. 86, pp. 319–328, Sep. 2018.

[11] M. Vinel, I. Ryazanov, D. Botov, and I. Nikolaev, “Experimental comparison

of unsupervised approaches in the task of separating specializations within professions in job vacancies,” in *Proc. Conf. Artif. Intell. Natural Lang.*, Cham, Switzerland: Springer, 2019, pp. 99–112.

[12] E. Malherbe, M. Cataldi, and A. Ballatore, “Bringing order to the job market: Efficient job offer categorization in E-recruitment,” in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 1101–1104.

[13] F. Saberi-Movahed, M. Rostami, K. Berahmand, S. Karami, P. Tiwari, M. Oussalah, and S. S. Band, “Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection,” *Knowl.-Based Syst.*, vol. 256, Nov. 2022, Art. no. 109884.

[14] I. Khaouja, I. Rahhal, M. Elouali, G. Mezzour, I. Kassou, and K. M. Carley, “Analyzing the needs of the offshore sector in Morocco by mining job ads,” in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1380–1388.

[15] R. Bekkerman and M. Gavish, “High-precision phrase-based document classification on a modern scale,” in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 231–239.