TWO-STAGE MACHINE LEARNING FRAMEWORK FOR ACCURATE JOB TITLE IDENTIFICATION IN ONLINE JOB ADVERTISEMENTS

[1]L. Priyanka, [2]Akula Saishivani
[1]Assistant Professor, [2]MCA Student
Department Of MCA Student
Sree Chaitanya College of Engineering, Karimnagar

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. In order to handle the situation of limited 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 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

The widespread use of the Internet in many industries due to the digitization of processes and the development of social media has resulted in a large amount of data that needs to be processed and analyzed quickly and efficiently to extract valuable insights that can help in decision-making [1]. In this context, data science techniques can be powerful tools for extracting information from large datasets, facilitating the process of classifying different types of data (e.g., text, images, and video) [2], and can also solve many other tasks that are handled in a traditional manner, which is often time and resource consuming.

Similarly, the job market shifted from traditional channels to online websites and job portals. This is because employers and recruiters share various job advertisements across different platforms to expand their reach and target more job seekers. This shift represents an opportunity to understand the needs of the job market from the vast amount of data shared daily, which can benefit many stakeholders [3]. In particular, identifying the requirements in terms of skills and occupations can help labor market analysts and policymakers foster employment and also help job seekers and students find suitable jobs and the training needed to successfully transition to the job market [4].

Classifying online job advertisements (ads) is not a trivial task. Indeed, the information contained in a job ad is expressed in plain text in a non-structured or semi-structured format, and the lexicon used by employers in the text is often very different from the occupational classifiers and the databases developed by human resources experts. In addition, job ads often include overly generic information that is not relevant to the position. This adds noise to the process of matching the job advertisement to its corresponding occupation. For instance, a job advertisement may have a title that includes information about the city where the job is located or some salary information. Also, the description can contain information about the company and information about other tasks that do not necessarily relate to the desired occupation. It is therefore necessary to apply advanced techniques for word and document representation and to use novel feature extraction methods to address these challenges.

Most of the proposed methods for dealing with occupation normalization consider it to be a classification or clustering problem. In this context, several text classifiers, ranging from traditional machine learning (ML) models to deep learning models, have been proposed for this task, such as support vector machine (SVM) [1], naïve bayes [5], k-nearestneighbor (KNN) [1], [5], artificial neural networks (ANNs) [6] and Bidirectional Encoder Representations from Transformers (BERT) [7]. While some studies used the combination of the title and the description to perform classification, the authors in [8] used only the title and found that 30% of the job offer titles did not contain enough information to identify the occupation. Similarly, the authors in [9] looked at the text of the job description only and found that each job description could correspond to more than one occupation. To the best of our knowledge, no previous study has examined the degree to which the title and description contribute when normalizing job ads. Classifying job ads using an occupational classifier or internal taxonomy has generally achieved satisfying results [6], [10]. However, these methodologies require human-labeled datasets with hundreds of thousands of examples, which is time-consuming and resource-intensive. In addition, updating the occupation description or including a newly created occupation in the occupational classifier is very difficult, as the entire training process

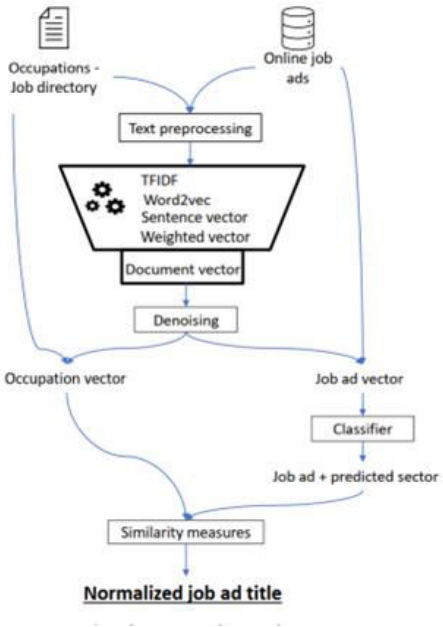
must be repeated. Also, most prior work only focuses on English-language job ads and uses specific occupational classifiers such as the Occupational Information Network (O*NET); the extension of existing work to job ads written in other languages is challenging. which makes it extremely difficult to replicate their methodologies in other languages.

On the other hand, using unsupervised models to identify the occupation, such as clustering [11] and field similarity [12], avoids training the model with labeled data, which are not always available; this is particularly relevant since we are dealing with a large number of occupations. The majority of previous works have relied on simple techniques for word embedding such as Bag of Words (BOW) [1], [12] or Term Frequency Inverse Document Frequency (TFIDF) [11] to generate word embedding and have applied averaging methods to calculate the document embedding. However, these techniques are considered weak in capturing the semantic relationships between the words, especially when we are dealing with job ads written by multiple employers who are using different lexicons. Therefore, word embedding approaches and feature extraction techniques [13] need to be closely monitored to achieve the highest results, as state-of the- art techniques do not perform well in all cases [11].

In this paper, we propose a job title identification methodology based on self-supervised and unsupervised machine learning algorithms with minimal labeling and high accuracy that can be replicated on data from other countries to overcome the limitations mentioned above. The proposed methodology consists of two steps: the classification of job ads by sector and the matching of job ads with occupations belonging to the predicted sector. The step of job ads classification is done using several text classifiers such as SVM, Naïve Bayes, Logistic Regression, and BERT to classify job ads into their corresponding sectors (e.g., Information Technology (IT), Agriculture) which will help us focus on the occupations of the predicted sector instead of using all the occupations from the occupational classifier. For the job title identification step, we compare different techniques for vector representation of texts and use several combinations and parameters to propose a customized document embedding strategy. We also test several feature selection methods to extract important keywords from the description and analyze the degree of contribution of the title and the description in improving the results. Finally, we calculate the similarity between the job ad representation and the occupation representations belonging to the predicted sector to choose the closest one. To do this, we collect the French occupational classifier ‘‘Pole Emploi’’ and about two hundred thousand job ads from job portals. When used to identify the occupation title on a random sample of job ads, our methodology achieves an overall accuracy of 76.5% and more than 85% for some sectors which is considered high accuracy compared to prior work. Furthermore, the effectiveness of our approach was validated with the help of a

team of domain experts who manually labeled a sample of our dataset. Finally, we applied our methodology to a dataset of 248,059 job ads in the French language to get an overview of the Moroccan job market, especially the IT sector. This study allows us to shed light on key sectors and occupations in the Moroccan job market where there is a high demand for IT profiles and Telemarketers which was identified by a previous study on the offshore sector in Morocco [14]. Using this methodology, we can identify emerging occupations that can help decision-makers including universities to take appropriate measures to adapt their programs and curricula, and to also help job seekers and students in their orientation by taking a career path that leads to employment [4].

2. SYSTEM ARCHITECTURE



3. EXISTING SYSTEM

Many studies have attempted to normalize job ads titles as a first step in structuring job ads before identifying the required skills based on job roles. An occupation is defined as a grouping of jobs that involve similar tasks and that require a similar skill set. It is important to note that occupations should not be confused with jobs or job titles. While a job is tied to a specific work context and executed by one person, occupations group jobs based on common characteristics. Identifying the required occupations in the job market can be considered as a top-down approach to discovering the required skills by inferring skills from structured skill bases that encompass full occupation descriptions such as the International Standard Classification of Occupations (ISCO) or O*NET.

There are two approaches to identifying job titles from job advertisements. The first approach uses supervised models to classify job titles, while the second approach uses unsupervised models to find the closest job title. In this section, we review the previous studies on job ad classification methods Many studies framed the task of job title identification as a text classification task where job ads were classified to their corresponding occupation based on the standard referential using SVM and KNN. In particular, in [1] and [18], CarrerBuilder. com used a multi-stage classifier to tackle a large number of classes which is almost similar to the application domain (online recruitment) used by LinkedIn's job title classification system [15], where they utilize a heavily manual phrase-based classification system dependent on short-text and a heavy reliance on crowd-sourced labeling of training samples. Moreover, in [16], they

leveraged string similarity, where similar job titles were fed to the siamese network to learn to classify job titles. For this task, they used an in-house taxonomy to classify the job titles instead of using O*NET and ISCO bases. Also in [6], [7], [8], and [10] they used text classifiers, from traditional machine learning models to deep learning models respectively based on ISCO occupation classifiers or on customized lists of occupations. Text classifiers used in [6] and [10] showed good performance in extracting the needed skills of some occupations, while text classifiers of [8] achieved a less interesting accuracy because they used only the title of the job ad and didn't include the description. Finally, the authors identified that about 30% of the job offer titles do not carry enough information to identify the occupation.

A similar study was described in [5] where the authors used a dataset from Kaggle to classify job titles based on the query description into 30 distinct classes corresponding to the top 30 occupations. They used several algorithms such as Bernoulli's Naïve Bayes, Multinomial Naïve Bayes, Random Forest, and Linear SVM and found that Linear SVM gives the best results for job title classification and that increasing the training set improves the accuracy. Finally, in [9] authors propose a multi-label classification approach for predicting relevant job titles from job description texts and consider that each job description may correspond to more than one occupation. They implement the algorithm presented in [7] using different pre-trained language models to apply it to the job titles prediction problem. They found that BERT with a multilingual pre-trained model obtained the highest result on their dataset and that the description alone is not enough for the prediction, so they need to reference extra information such as job name, job level, and job requirements.

Disadvantages

- The main disadvantage of text classifiers is the expense of data acquisition for training with many thousands of groups of occupations, often not too dissimilar from one another.
- In an existing system, due to the lack of labeled datasets that can be used for the training step, we opted to use a combination of the two approaches.

4. PROPOSED SYSTEM

In this paper, we propose a job title identification methodology based on self-supervised and unsupervised machine learning algorithms with minimal labeling and high accuracy that can be replicated on data from other countries to overcome the limitations mentioned above. The proposed methodology consists of two steps: the classification of job ads by sector and the matching of job ads with occupations belonging to the predicted sector.

The step of job ads classification is done using several text classifiers such as SVM, Naïve Bayes, Logistic Regression, and BERT to classify job ads into their corresponding sectors (e.g., Information Technology (IT), Agriculture) which will help us focus on the occupations of the

predicted sector instead of using all the occupations from the occupational classifier. For the job title identification step, we compare different techniques for vector representation of texts and use several combinations and parameters to propose a customized document embedding strategy. We also test several feature selection methods to extract important keywords from the description and analyze the degree of contribution of the title and the description in improving the results.

Finally, we calculate the similarity between the job ad representation and the occupation representations belonging to the predicted sector to choose the closest one. To do this, we collect the French occupational classifier ‘‘Pole Emploi’’ and about two hundred thousand job ads from job portals. When used to identify the occupation title on a random sample of job ads, our methodology achieves an overall accuracy of 76.5% and more than 85% for some sectors which is considered high accuracy compared to prior work.

Furthermore, the effectiveness of our approach was validated with the help of a team of domain experts who manually labeled a sample of our dataset. Finally, we applied our methodology to a dataset of 248,059 job ads in the French language to get an overview of the Moroccan job market, especially the IT sector. This study allows us to shed light on key sectors and occupations in the Moroccan job market where there is a high demand for IT profiles and Telemarketers which was identified by a previous study on the offshore sector in Morocco [14]. Using this methodology, we can identify emerging occupations that can help decision-makers including universities to take appropriate measures to adapt their programs and curricula, and to also help job seekers and students in their orientation by taking a career path that leads to employment [4].

Advantages

• We propose a methodology for occupation identification in a scenario of a lack of labeled data so that it can be replicated for other languages and countries.

• We provide a comparison of document representation strategies for solving the problem of occupation identification and identify the degree of contribution of the title and description of the job ad in the matching process.

• We draw insights on the Moroccan IT job market needs in terms of occupation and construct a Moroccan job ads dataset in the French language, which can relieve the limitation in this field.

5. IMPLEMENTATION

Modules description

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse Datasets and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Job Title Identification Type, View Job Title Identification Type Ratio, Download Predicted Data Sets, View

Job Title Identification Type Ratio Results, View All Remote Users.

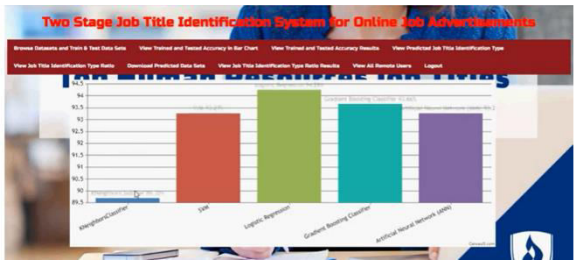
View and Authorize Users

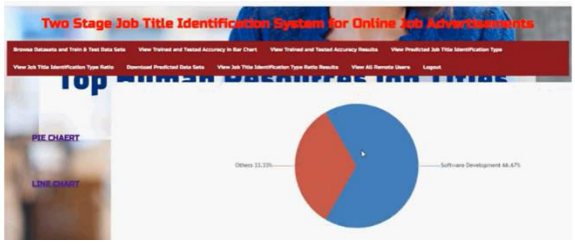
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, Predict Job Title Identification Type, VIEW YOUR PROFILE.

6. RESULTS





7. CONCLUSION

Our two-stage job tile identification approach, which is based on semi-supervised and unsupervised machine learning algorithms with little labelling, is presented in this work. 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 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.

[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 unslupervised 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. ACMSIGIR 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.