A Two-Step Procedure for Detecting Posting Titles in Online Adverts

MRS. S.LAVANYA [1], MISS.K.KAVYA [2]

#1 Assistant Professor in the department of IT at DVR & DR HS MIC COLLEGE OF TECHNOLOGY (Autonomous), Kanchikacherla (NTR Dist, AP)

#2 MCA Student In The Department Of Computer Applications at DVR & Dr HS MIC College of Technology (Autonomous), Kanchikacherla, NTR District, A.P

ABSTRACT— Data science approaches are very useful for mining massive databases for insights. There has been a lot of recent focus on analysing the employment market via the categorization of internet job adverts. In order to successfully determine the job title from an advertising, many multi-label classification methods have been developed, such as clustering and self-supervised learning. These methods, however, are only applicable to databases tailored to the American labor market, such as O*NET, and they need tagged datasets containing many thousands of samples. In this study, we tackle the issue of tiny datasets by presenting a two-stage algorithm for job title identification. To begin, we sort the job postings by industry using Bidirectional Encoder Representations from Transformers (BERT)—for example, IT jobs and agricultural jobs. The next step is to locate the most closely related job title from the anticipated sector's list of jobs using unsupervised machine learning methods and certain similarity measurements. To solve the problems of processing and categorizing job advertising, we also suggest a new method of document embedding. Based on our experiments, the suggested two-stage method increases the accuracy of job title detection by 14%, reaching over 85% in some industries. In addition, when comparing methods based on the Bag of words model to those that use document embedding-based techniques such weighting schemes and noise reduction, we discovered that the classification accuracy is improved by 23.5%. Further tests confirm that the suggested approach is superior than or on par with the state-of-the-art approaches. New and in-demand jobs in Morocco have been discovered by applying the suggested technique to data from the country's labour market.

INTRODUCTION

A great deal of data has to be processed and evaluated quickly and effectively to derive useful insights that may aid in decision-

making. This is because the Internet has become ubiquitous in many sectors as a consequence of process digitalization and the growth of social media. In this setting, data science methods can be useful for many things that are currently done in a more conventional way, which can be a waste of time and resources. For example, they can help with data classification (e.g., text, images, and video) and information extraction from big datasets. Websites and job portals also replaced more conventional methods of advertising for employment. This is due to the fact that in order to reach a wider audience of potential candidates, recruiters and businesses post job ads on many sites. Many parties stand to gain from this change, since it opens the door to gleaning insights about the labour market from the mountains of data that are exchanged every day. In instance, determining the necessary skills and vocations may aid policymakers and labour market analysts in creating jobs, while also assisting students and job-seekers in locating appropriate training and appropriate career opportunities. It is not a simple effort to categorize internet job adverts. Employers' language in job ads differs significantly from occupational classifiers and databases created by HR professionals, and the information included in these ads is conveyed in plain English in a non-structured or semi-structured fashion. On top of that, it's not uncommon for job postings to include irrelevant, general material. This complicates the task of determining which profession is most closely related to the job posting. As an example, the title of a job posting could include details like the location in which the position is situated or even pay information. In addition to details about the desired job, the description might also include details about the firm and other duties. To solve these problems, new feature extraction approaches and sophisticated word and text representation techniques are required. Occupation normalization is treated as a clustering or classification issue in most of the offered solutions. Various text classifiers, including naïve bayes, support vector machine (SVM), artificial neural networks (ANNs), k-nearest neighbour (KNN), ANNs, and Bidirectional Encoder Representations from Transformers (BERT), have been suggested for this task within this context. These models range from traditional machine learning (ML) to deep learning. Despite the fact that other studies have used both the title and the description to classify job offers, the authors of relied just on the title and discovered that thirty percent of the titles lacked sufficient information to designate

RELATED WORK

Carotene: A Job Title Classification System for the Online Recruitment Domain

In the online job recruitment domain, accurate classification of jobs and resumes to occupation categories is important for matching job seekers with relevant jobs. An example of such a job title classification system is an automatic text document classification system that utilizes machine learning. Machine learning-based document classification techniques for images, text and related entities have been well researched in academia and have also been successfully applied in many industrial settings. In this paper we present Carotene, a machine learning-based semi-supervised job title classification system that is currently in production at CareerBuilder. Carotene leverages a varied collection of classification and clustering tools and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in a cascade classifier architecture. We first present the architecture of Carotene, which consists of a two-stage coarse and fine level classifier cascade. We compare Carotene to an early version that was based on a flat classifier architecture and also compare and contrast Carotene with a third party occupation classification system. The paper concludes by presenting experimental results on real world industrial data using both machine learning metrics and actual user experience surveys.

ScienceDirect Web Data Extraction Approach for Deep Web using WEIDJ

Data extraction is one of the most prominent areas in data mining analysis that is been extensively studied especially in the field of data requirements and reservoir. The main aim of data extraction with regards to semi-structured data is to retrieve beneficial information from the World Wide Web.

A Hybrid Approach to Managing Job Offers And Candidates

The evolution of the job market has resulted in traditional methods of recruitment becoming insufficient. As it is now necessary to handle volumes of information (mostly in the form of free text) that are impossible to process manually, an analysis and assisted categorization are essential to address this issue. In this paper, we present a combination of the E-Gen and Cortex systems.

Education Path: Student orientation based on the job market needs

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

Using Machine Learning for Labour Market Intelligence

It can be used to build a language-independent knowledge base for analysis purposes instead of only matching CVs with job vacancies. This could be a new approach for natural language texts (Boselli et al., 2017). This approach will help with deep data analysis and enable data science to aggregate and group data to get meaningful information and build a dashboard for the decisionmakers

Challenge: Processing web texts for classifying job offers

Today the Web represents a rich source of labour market data for both public and private operators, as a growing number of job offers are advertised through Web portals and services. In this paper we apply and compare several techniques, namely explicit-rules, machine learning, and LDAbased algorithms to classify a real dataset of Web job offers collected from 12 heterogeneous sources against a standard classification system of occupations.

METHODOLOGY

Calculate metrics: Using this module, defining function to calculate accuracy and other metrics

Train Logistic Regression: Using this module, training logistic regression on TFIDF features and it got 84% accuracy

Train SVM: Using this module, training SVM on TFIDF features and it got 83% accuracy

Train Naïve Bayes: Using this module, training Naïve Bayes on TFIDF features and it got 51% accuracy

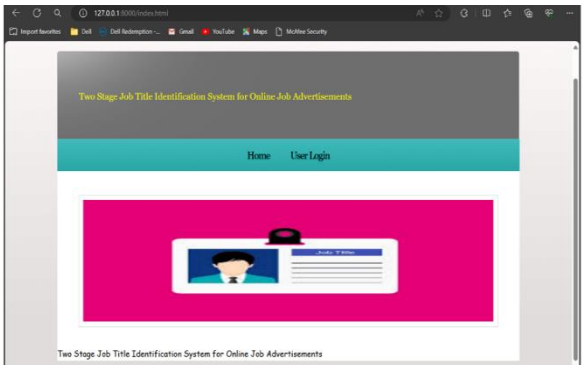
Train CNN: Using this module, training CNN on TFIDF features and it got 96% accuracy

Train BERT: Using this module, training BERT on TFIDF features and it got 88% accuracy

Comparison Graph: Using this module, displaying all algorithm performance where x-axis represents algorithm names and y-axis represents accuracy and other metrics in differnet colour bars

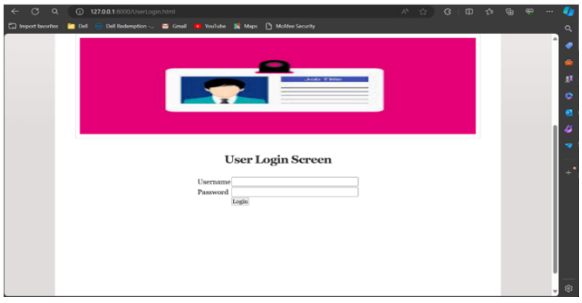
Predict: Using this module, reading JOB description from TEST data and then predicting JOB TITLE

RESULT AND DISCUSSION

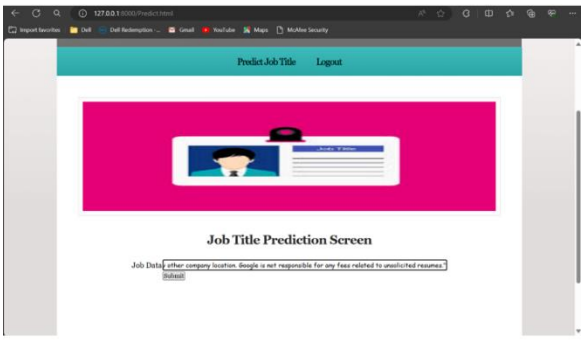


In above screen displaying the home page of

Two stage job title identification system for online job advertisements



In above screen displaying User login page .In this user can login with their credentials .If it is matched then user can successful login their page.



After successful login they can find a predict job title in that they can add data by uploading a dataset

In the above screen in first line,we can see JOB Description and then after ➔ arrow symbol can see predicted JOB title as big data Engineer.

CONCLUSION

In this paper, we present a two-stage job tile identification methodology based on semi-supervised and unsupervised machine learning algorithms with minimal labelling. In particular, for each job ad based on similarity measures, we find the most

appropriate occupation using a standard occupational classifier. During the conducted experiments and after pre-processing the collected job ads, we tested several word and document representation methods such as TFIDF, neural language models that rely on distributional semantics (Word2Vec, Fast Text), and deep contextualized word representation (BERT). They were all subjected to several weighting strategies in order to reduce the impact of irrelevant words, especially in the description. Then, we tested various balance factors to identify the degree of contribution of both the title and the description to the process. According to the experiment results, classifying the job ads by sector improved the accuracy of our methodology by 14% since the similarity measures between the job ad and the occupations will be applied only within the predicted sector instead of using all the occupations from the referential. For document representation, we found that results using W2V outperformed BERT since there is a difference in vocabulary between the training dataset and job vacancies. However, in the case where the sector is not specified, we found that BERT provides the most accurate results. When it comes to weighting strategies, results show that uniform and frequency word weighting work best for short text (job ad titles, occupation titles), as these are not sensitive to word weighting, while the TFIDF weighting strategy for long text (job ad descriptions, occupation descriptions) significantly improves performance. In addition, we found that document embedding using only the top N selective words from the description using weighting scores gives the most accurate results among all the configurations we tested since we add relevant context to the title. Finally, experiments also verify the effectiveness of using both the title and the description in the matching process. They also verify that we should not give them equal weights because the title is more relevant since it contains more dense words related to the job. These findings helped us improve the accuracy of our methodology by 34% over the baseline. Our results - in terms of performance – are comparable to those obtained by the classification approach. Specifically, we obtained an overall accuracy of 76.5%, which can sometimes exceed 85% depending on the sector, such as the health sector and hotel & tourism sector. Furthermore, these findings can also be applied to improve the accuracy of the classifier when considering the task of job title identification as a classification problem. Finally, this methodology can be replicated in

other languages using other occupation classifiers with minimal interaction to normalize the job ads and get insights from them. The proposed technique has been tested in a real-life setting framed within the project called ‘‘Data science for improved education and employment in Morocco’’ supported by USAID which aims at analyzing the job market needs and extracting skills from them [4]. It can also be applied in the process of defining training courses by universities based on job market needs. At the same time, youth and job seekers looking for employment can benefit from the results of studies using this methodology to analyze the labor market. In the future, we intend to add a step of job enrichment with skills terms based on the occupation description so that the job ad and occupation description are as similar as possible because recruiters do not follow a specific format when writing job advertisements. We also intend to do more cleaning of the list of top N words generated by weighting strategies to keep only relevant words. Furthermore, we plan to train our own Word2Vec model on sentences related to jobs in French, which may increase the accuracy of our methodology.

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), Y. Altun, K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Zitnik, M. Ceci, and S. Dzeroski, Eds. Cham, Switzerland: Springer, 2017, pp. 330–342.

[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 multilabel 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. 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.

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.

[14] 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.

[15] P. Neculoiu, M. Versteegh, and M. Rotaru, ‘‘Learning text similarity with siamese recurrent networks,’’ in Proc. 1st Workshop Represent. Learn. (NLP). Berlin, Germany: Association for Computational Linguistics, 2016, pp. 148–157. Accessed: Feb. 22, 2023. [Online]. Available: http://aclweb.org/anthology/W16-1617, doi: 10.18653/v1/W16-1617.

[16] I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Woon, ‘‘Data mining approach to monitoring the requirements of the job market: A case study,’’ Inf. Syst., vol. 65, pp. 1–6, Apr. 2017.

[17] Y. Zhu, F. Javed, and O. Ozturk, ‘‘Document embedding strategies for job title classification,’’ in Proc. 30th Int. Flairs Conf., 2017, pp. 55–65. Accessed: Oct. 4, 2022. [Online]. Available: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15470

[18] AUTHOR PROFILES

Mrs. S.LAVANYA completed her Bachelor of Technology in Computer Science and Engineering. She completed her Masters of Technology in Computer Science and Engineering from JNTU KAKINADA UNIVERSITY. Currently working as an Assistant Professor in the department of IT at DVR & DR .HS MIC COLLEGE OF TECHNOLOGY (Autonomous),Kanchikacherla (NTR Dist, AP). Her areas of interest are Data Mining, Cloud Computing and Machine Learning & Networks.

Miss. KAVYA KARANAM, as MCA student in the department of DCA at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District. She has completed Bcom(GENERAL) in Sri Krishnaveni Mahila College. Her areas of interests are Networks, Machine Learning ,java and Cloud Computing.