10.48047/jocaaa.2024.32.01.25

System for Job Title Identification for Online Job Advertisements Using CNN2D Algorithm

Mr.Bibhu Ranjan Sahoo[1], Y,Aarthi [2], H.Priyanka [3], A.Manisha[4]
[1] Associate Professor, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India
Email: bibhusahoo02@gmail.com
[2,3,4,]B.Tech Student, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

ABSTRACT:

Large datasets may be mined for information using effective data science approaches. Classifying online job adverts has gained a lot of interest lately as a means of analysing the employment market. A number of multi-label classification techniques (such as clustering and self-supervised learning) have been developed to successfully determine the profession from a job advertising. Nevertheless, these methods rely on labelled datasets including hundreds of thousands of samples and concentrate on certain databases like the Occupational Information System (O*NET) that are more tailored to the US labour market. To address the problem of limited datasets, we describe in this work an a two-phase job title identification algorithm. First, we categorise the job advertising by industry (e.g., agriculture, information technology) using representations of bidirectional encoders from Transducers (BERT). Next, we identify the job title that most closely matches the list of jobs inside the anticipated sector using unsupervised machine learning methods and a few similarity metrics. In order to solve the problems of processing and categorising employment adverts, we also suggest a unique document embedding technique. Our test findings demonstrate that the suggested two-stage method increases job title recognition accuracy by 14%, reaching over 85% in some industries. Furthermore, we discovered that the classification accuracy is increased by 23.5% when document embedding-based techniques such noise reduction and weighting methods are used in place of approaches that rely on the Bag of words model. Additional assessments confirm that the suggested technique performs on par with or better than the cutting-edge approaches. The suggested technique has been used to discover high-

10.48047/jocaaa.2024.32.01.25

demand and emerging jobs in Morocco by analysing employment market data.

Keywords: BERT, job market analysis, job title classification, job title identification, machine learning, natural language processing.

INTRODUCTION

Due to the digitalization of processes & the growth of social media, the Internet is now widely used in many sectors, which has led to the accumulation of a lot of data that has to be processed or analysed fast in order to extract insightful information that may aid in decision-making. Within this framework, data science methods can be effective instruments for drawing insights from sizable databases, streamlining the categorization of various data kinds (text, photos, and video, for example), and resolving numerous other issues that are addressed through conventional methods, which are frequently laborious and resource-intensive. In a similar vein, internet job portals and websites have replaced conventional routes in the employment market. This is a result of recruiters and businesses disseminating diverse job ads on many channels in an effort to increase their reach and attract more applicants. This change offers many stakeholders the chance to gain insight into the demands of the labour market by using the massive volume of data that is provided on a regular basis. Specifically, determining the skills and occupation needs may assist labour market economists and policymakers in promoting employment, as well as assist job seekers as well as learners in locating appropriate positions and the training necessary to effectively enter the workforce [4]. Sorting through job adverts on the internet is a difficult chore. Indeed, a job advertisement's content is provided in plain English in a semi-structured or unstructured fashion, and employers' use of vocabulary in the ad often differs greatly from occupational classifiers or databases created by human resources specialists. Furthermore, too general material unrelated to the role is often included in job advertisements. This complicates the process of connecting the employment posting to the appropriate profession. For example, the title of a job posting could include details like the location in which the position is situated or the wage range. Furthermore, the job description may include details about the organisation as well as additional responsibilities unrelated to the intended career path. To overcome these obstacles, it is consequently required to use

10.48047/jocaaa.2024.32.01.25

cutting-edge strategies for word or document representation as well as cutting-edge feature extraction algorithms. The majority of strategies that have been put out to address occupation normalisation see the issue as one of categorization or grouping. Support vector machine, or SVM, naive bayes, k-nearest neighbour (KNN), ANNs (artificial neural networks) and bidirectional representations of encoders from Transformers (BERT) are a few of the text classifiers that have been proposed for this task, ranging from traditional models based on machine learning (ML) to deep learning models. While other studies classified using both the title or the description, the authors in [8] discovered that 30% of job offer titles were insufficient to identify the profession when they utilised simply the title. Similar findings were made by the authors in, who merely examined the job description's language and discovered that every position description might apply to many occupations. To the best of our knowledge, no prior research has looked at how much the job description and title contribute to the normalisation of job adverts. Using an internal taxonomy or occupational classifier to categorise job adverts has typically produced satisfactory results. However, the human-labelled datasets containing tens of millions of instances required by these approaches are resource- and time-intensive. Additionally, since the whole training process has to be performed, it is exceedingly challenging to update the occupation description or add a newly generated occupation to the occupational classifier. It is also difficult to extend previous work to job advertising written in other languages since it primarily focused on English-language job ads and used specialised occupational classifiers like the ones used by the Occupational Information Network (O*NET). It makes it very challenging to translate their methods into other languages. However, given we are working with a large variety of professions, it is especially important to avoid training the model with labelled data by employing unsupervised methods, such clustering and domain similarity, to identify the occupation. Most prior publications have used averaging approaches to compute the document embedding and have relied on basic word embedding techniques like Bag of Words (BOW), or Term Frequency Inverted Document Frequency (TFIDF) to create word embedding. These methods, however, are thought to be inadequate for capturing the semantic links between the words, particularly when dealing with job advertisements prepared by many employers

10.48047/jocaaa.2024.32.01.25

using various lexicons. Since state-of-the-art methods are not always effective, word embedding methods or feature extraction techniques must be constantly watched in order to get the best results. To solve the aforementioned restrictions, we provide in this work a job title recognition approach based on self-supervised with unsupervised machine learning algorithms that have poor labelling and high accuracy that can be repeated on data from various nations. The two stages of the suggested technique are to classify job advertising based on industry and then match job ads with professions that fall within the identified industry. Several text classifiers are used in the job ads classification step, including SVM, Naïve Bayes, Logistic Regression, and BERT. These classifiers help us focus on the occupations within the predicted sector, rather than using all the occupations from the occupational classifier, and help classify job ads into the appropriate sectors (e.g., Information Technology (IT), Agriculture). In order to suggest a customised document embedding approach for the job title identification stage, we examine several methods for vector representations of texts using a variety of combinations of parameters. Additionally, we evaluate several feature selection techniques to extract significant keywords from the summary and assess the relative contributions of the title and description to better outcomes. In order to choose the closest representation, we lastly compute the similarity among the job ad representation and the job representations that correspond to the anticipated sector. We gather the French occupational classifier "Pole Employ" and over 200,000 job adverts from employment sites in order to do this. Our technique obtains an overall accuracy of 76.5% as well as greater than 85% for specific sectors when used to identify the profession title on an arbitrary number of job advertising, which is deemed high accuracy relative to previous studies. Additionally, our method's efficacy was confirmed by a group of subject-matter specialists who individually annotated a portion of our dataset. Finally, in order to get a general understanding of the Moroccan employment market, particularly in the IT sector, we used our technique to a dataset of 248,059 job advertising in the French language. This research enables us to throw light on important industries and professions in the Moroccan labour market, where telemarketers and IT profiles are in great demand, as noted by an earlier investigation on Morocco's offshore industry. This technique allows us to find new professions that may assist policymakers,

10.48047/jocaaa.2024.32.01.25

academic institutions, and job seekers in orienting themselves by choosing a professional path and leads to employment, as well as universities in making the necessary changes to their programmes and curriculum.

RELATED WORK

Carotene: A System of Job Title Categorization for Online Employment

For the purpose of connecting job seekers with appropriate positions, precise job classification and resume categorization to occupational groups is crucial in the web-based job recruiting space. An automated machine learning-based text-based file classification system is one example of this kind of job title classification system. Academic research on machine learning-based document categorization methods for text, pictures, and related items has been extensive, and these methods have found widespread use in industry. In this work, we introduce Carotene, a semi-supervised job title categorization system that uses machine learning and is now being used at CareerBuilder. Using a wide range or classification and clustering methods and approaches, Carotene addresses the difficulties involved in creating an extensible system of categorization for a vast taxonomy of job types. It uses a cascade classifier design to include these methods. First, we introduce the two-stage coarse and finer level classifier cascade that is the foundation of the Carotene design. We also contrast and compare Carotene with a third-party occupation classification system, as well as with an early version of the system that was built on a flat classifier architecture. The experimental findings on real-world industrial data utilising both real user experience surveys and machine learning measures are shown in the paper's conclusion.

ScienceDirect WEIDJ-Based Web Data Extraction Method for the Deep Web

In data mining analysis, data extraction constitutes one of the most well-known topics that has been thoroughly investigated, particularly in the areas of data needs and reservoir. Retrieving useful information from the Internet is the primary goal of data extraction when it comes to semi-structured data. Although it cannot be done by any search engine, retrieving material from the deep web, commonly referred to as the vast

10.48047/jocaaa.2024.32.01.25

web, needs submitting a form. The different structure of web sites makes data mining systems and robotic data extraction highly laborious. The majority of earlier data extraction methods dealt with other kinds of data, including text, audio, video, and so on; nevertheless, there are still not enough studies that concentrate on utilising images as data. The DOM, or Document Object Model, represents the cutting edge of data extraction methods associated with image data mining research. The technique used to address partially structured information extracted from the web was DOM. Unfortunately, it has been discovered that noisy information and long processing times are problems that arise as HTML texts become bigger. In this study, we take into account the extraction of web data from the deep web and, in response to the encouraging outcomes of mining a larger amount of online data from a variety of image formats, we suggest an enhanced model called Wrapped Extraction of Image employing DOM and JSON (WEIDJ). In order to evaluate the effectiveness of the suggested model, we contrast its data extraction performance by page extraction level with that of other approaches, including VIBS, MDR, DEPTA, and VIDE. It produced the greatest results in terms of F-measure (98.9547), Precision (100), and Recall (97.93103).

A combination of methods for handling job offers and applicants

Traditional techniques of recruiting are no longer adequate due to the changes in the labour market. These days, handling amounts of data (mainly free language) that are too big to manage by hand makes sense. To solve this problem, an analysis and helped classification are crucial. We describe an E-Gen with Cortex system combo in this study. E-Gen endeavours to do analysis and classification of employment offers in conjunction with the applicants' answers. To address the issue of profiling applications in accordance with a particular job offer, the E-Gen system technique is based on vectorial or probabilistic models. A system for statistical automated summarization is called Cortex. E-Gen employs Cortex as an effective filter in this study to remove superfluous information from candidate responses. Our primary aim is creating a system to support a recruiting consultant, and the outcomes of the suggested combination outperform those of E-Gen operating alone in this regard.

10.48047/jocaaa.2024.32.01.25

Using Machine Learning to Classify Online Job Advertisements

Real-time labour market tracking is greatly facilitated by the increasing use of the Internet to post job advertisements. This is the goal of the discipline known as labour market insight (LMI), which is becoming more and more important in the creation and assessment of EU labour market policy. In fact, compared to traditional survey-based analyses, the analysis of online job openings gives labour market participants a competitive edge as it shortens the analysis's time to market by shifting towards a fact-based decision-making paradigm. This article describes our method for automatically categorising millions of job listings on the Web according to a common occupational taxonomy. We demonstrate how machine learning has been used to formulate this issue in the context of text categorization. We also describe the advantages to end users and demonstrate our approach's application to a few real-world applications.

METHODOLOGY

So, author of this paper employing two stage Job Title Identification and these stages include

1) In first Stage author using Bidirectional Encoder Representations from Transformers (BERT) to first classify the job ads according to their corresponding sector (e.g., Information Technology, Agriculture). BERT is used to convert unstructured text data into numeric vector by considering semantic similarity.
2) In stage two author employing Euclidean Distance algorithm to measure similarity between train and test data to find closest matching Job Title. This similarity measure will work with small number of labels and doesn't require huge labels of data.

RESULT AND DISCUSSION

In above screen importing require python classes and packages

10.48047/jocaaa.2024.32.01.25

In above screen displaying all algorithm performance where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars

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

CONCLUSION

This study presents a two-stage technique for job tile recognition based on minimally labelled machine learning algorithms, both semi-supervised and unsupervised. Specifically, we use a typical occupational classifier to choose the best profession for each job post based on similarity measurements. After pre-processing the gathered job advertising, we examined a number of word or document representation techniques throughout the conducted tests, including deep contextualised word representation (BERT), neural models of language who rely on distributional semantic (Word2Vec, Fast Text), and TFIDF. To lessen the effect of unnecessary words, particularly in the description, they were all put through a number of weighing algorithms. The degree to which the title and description contributed to the process was then determined by testing different balancing variables. As the indicators of similarity between the job posting and the occupations would only be used inside the projected sector rather than utilising all the occupations used as reference, categorising the job advertising by sector enhanced our methodology's accuracy by 14%, according to our experiment findings. Because there is a lexical variation between the training set and job openings, we discovered that W2V produced better results for document representation than BERT. We discovered that BERT yields the best accurate answers when the sector is left unspecified. In terms of weighting strategies, the findings indicate that while long text (job ad descriptions,

10.48047/jocaaa.2024.32.01.25

occupation descriptions) is not sensitive to word weighting, uniform and frequency word weighting performs best for short text (job ad titles, occupation titles). On the other hand, TFIDF weighting significantly improves performance. Furthermore, by adding pertinent context to the title, we discovered that document embedding with just the top N selected words from the summary utilising weighting scores produces the best accurate results out of all the setups we examined. Experiments also confirm that utilising the title and description in the process of matching is beneficial. They further confirm that we should not assign them equal weights since the job title is more relevant due to its usage of more complex vocabulary. Thanks to these results, we were able to raise our methodology's accuracy by 34% over the baseline. Performance-wise, our outcomes are on par with those of the categorization strategy. In particular, we scored an overall accuracy of 76.5%, which, depending on the industry, may sometimes surpass 85%, as in the case of the hotel and tourist and health sectors. Moreover, these results may also be used to enhance the classifier's accuracy when the job title recognition task is regarded as a classification issue. In order to normalise the job advertising and get insights from them, this process may be easily duplicated in other languages with minimum intervention by using other occupation classifiers. Within the framework of the USAID-supported project "Data science to improve education and employment in Morocco," which seeks to analyse job market demands and extract skills from them, the suggested method has been evaluated in a real-world environment. It may also be used when colleges are creating their curricula according to the demands of the labour market. The findings of research that examine the labour market using this technique may also be advantageous to young people and job seekers. Since recruiters don't always follow a set structure when creating job advertising, we want to include a stage of job enrichment in the future that uses skills words based on the occupation description to make the job ad and occupation description as comparable as possible. Additionally, we want to further purify the top N word list produced using weighing algorithms in order to retain only relevant terms. Additionally, in order to improve the accuracy of these methods, we want to train our personal Word2Vec model on French phrases relating to occupations.

REFERENCES

10.48047/jocaaa.2024.32.01.25

[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 Compute. Service Appl., Mar. 2015, pp. 286–293.

[2] M. S. Pear, R. Rumsiyah, 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. Bechet, M. Roche, J.-M. Torres-Moreno, and M. El-Bezel, ‘‘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. Sibi, ‘‘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. mercurial, and M. Messianic, ‘‘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. Millikanian, D. Malerba, J. Stefanowski, J. Read, M. Zitnik, M. Ceci, and S. Demoski, 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. Compute. Common. Technol. (RIVF), Oct. 2020, pp. 1–6.

[8] F. Amato, R. Boselli, M. Cesarini, F. mercurial, M. Messianic, V. Moscato, F. Persia, and A. Picariello, ‘‘Challenge: Processing web texts for classifying job offers,’’ in Proc. IEEE 9th Int. Conf. Semantic Compute. (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. Compute. Sci. (NICS), Dec. 2021, pp. 513–518.

[10] R. Boselli, M. Cesarini, F. mercurial, and M. Messianic, ‘‘Classifying online job

10.48047/jocaaa.2024.32.01.25

advertisements through machine learning,'' Future Gener. Compute. Syst., vol. 86, pp. 319–328, Sep. 2018.

[11] M. Vanel, I. Ryazanov, D. Butov, and I. Nikolaev, ''Experimental comparison of unsupervised approaches in the task of separating specializations within professions in job vacancies,'' in Proc. Conf. Arif. Intel. 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. Retry., Aug. 2015, pp. 1101–1104.

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

[14] I. Khouja, I. Rahhal, M. Emomali, G. Mezz our, I. Kasson, and K. M. Carley, ''Analysing 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. Know. Discovery Data Mining, Aug. 2011, pp. 231–239.

[16] P. Nebulous, 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.

[17] I. Karakatsanis, W. AL Khader, F. McCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Won, ''Data mining approach to monitoring the requirements of the job market: A case study,'' Inf. Syst., vol. 65, pp. 1–6, Apr. 2017.

[18] 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://

10.48047/jocaaa.2024.32.01.25

www.aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15470

[19] F. Colace, M. D. Santo, M. Lombardi, F. Mercuria, M. Messianic, and F. Pascale, ‘‘Towards labour market intelligence through topic modelling,’’ in Proc. Annu. Hawaii Int. Conf. Syst. Sci., 2019, pp. 1–10.