Journal of Management & Entrepreneurship
ISSN 2229-5348
UGC Care Group I Journal
Vol-13 Issue-02 July 2024

DUAL STAGE JOB TITLE RECOGNITION SYSTEM FOR ONLINE JOB LISTINGS

Mr. B. SURESH REDDY[1], Ms. A. SIRISHA[2]

#1 Assistant. Professor in the department of MCA at QIS College of Engineering &Technology (Autonomous), Vengamukkapalem(V), Ongole, Prakasam dist.
#2 PG Scholar, Department of Master of Computer Applications, QIS College of Engineering & Technology, Vengamukkapalem, Ongole, Prakasam dist.

ABSTRACT

Data science techniques are powerful tools for extracting knowledge from large datasets. Analyzing the job market by classifying online job advertisements (ads) has recently received much attention. Various approaches for multi-label classification (e.g., self-supervised learning and clustering) have been developed to identify the occupation from a job advertisement and have achieved a satisfying performance. However, these approaches require labeled datasets with hundreds of thousands of examples and focus on specific databases such as the Occupational Information Network (O*NET) that are more adapted to the US job market. In this paper, we present a two-stage job title identification methodology to address the case of small datasets. We use Bidirectional Encoder Representations from Transformers (BERT) to first classify the job ads according to their corresponding sector (e.g., Information Technology, Agriculture). Then, we use unsupervised machine learning algorithms and some similarity measures to find the closest matching job title from the list of occupations within the predicted sector. We also propose a novel document embedding strategy to address the issues of processing and classifying job ads. Our experimental results show that the proposed two-stage approach improves the job title identification accuracy y 14% to achieve more than 85% in some sectors. Moreover, we found that incorporating document embedding-based approaches such as weighting strategies and noise removal improves the classification accuracy by 23.5% compared to approaches based on the Bag of words model. Further evaluations verify that the proposed methodology either outperforms or performs at least as well as the state-of-the-art methods. Applying the proposed methodology to Moroccan job market data has helped identify emerging and high-demand occupations in Morocco.

I.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 and 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 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 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 and they used text classifiers, from traditional machine learning models to deep learning models respectively based on ISCO occupation lassifiers or on customized lists of occupations. Text classifiers used in and slowed good performance in extracting the needed skills of some occupations, while text classifiers of 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 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 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 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

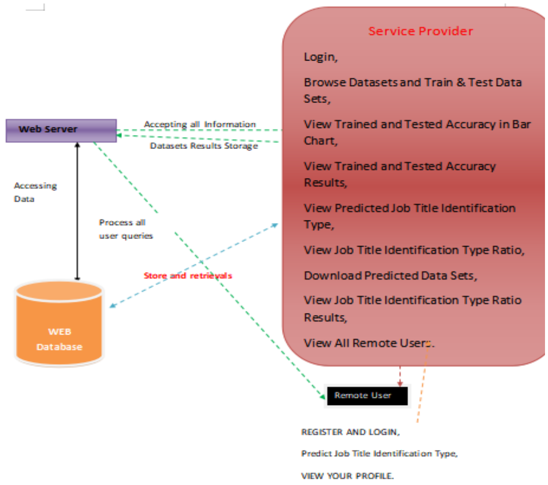
- The main disadvantage of text classifiers is the expense of data acquisition for training with many thousands of groups of occupations, often not too dissimilar from one another.
- In an existing system, due to the lack of labeled datasets that can be used for the training step, we opted to use a combination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

II.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 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 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.

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.



III.ALGORITHMS

1.K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Algorithm The k-nearest neighbors classification algorithm

Input:

D: a set of training samples $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$
k: the number of nearest neighbors
$d(\mathbf{x}, \mathbf{y})$: a distance metric
$\mathbf{x}$: a test sample

for each training sample $(\mathbf{x}_i, y_i) \in D$ **do**
  Compute $d(\mathbf{x}, \mathbf{x}_i)$, the distance between $\mathbf{x}$ and $\mathbf{x}_i$
Let $N \subseteq D$ be the set of training samples with the k smallest distances $d(\mathbf{x}, \mathbf{x}_i)$
return the majority label of the samples in N

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset.

2.SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

Algorithm 1: SVM

1. Set $Input = (x_i, y_i)$, where $i = 1, 2, \cdots, N, x_i = R^n$ and $y_i = \{+1, -1\}$.
2. Assign $f(X) = \omega^T x_i + b = \sum_{i=1}^{N} \omega^T x_i + b = 0$
3. Minimize the QP problem as, $min\ \varphi(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C.(\sum_{i=1}^{N} \xi_i)$.
4. Calculate the dual Lagrangian multipliers as $min\ L_P = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{N} x_i y_i(\omega x_i + b) + \sum_{i=1}^{N} x_i$.
5. Calculate the dual quadratic optimization (QP) problem as $max\ L_D = \sum_{i=1}^{N} x_i - \frac{1}{2}\sum_{i,j=1}^{N} x_i x_j\ y_i\ y_j\left(x_i, x_j\right)$.
6. Solve dual optimization problem as $\sum_{i=1}^{N} y_i x_i = 0$.
7. Output the classifier as $f(X) = sgn(\sum_{i=1}^{N} x_i y_i(x \cdot x_i) + j$.

3.Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar. Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Algorithm 2: PSEUDO code for logistic regression algorithm

```
Step1: Function grad (predictor_attributes, target_attribute, weights)
          {
            Calculate gradient_descent;
            Return weights + learning_rate * gradient_descent;
          }
Step2: Normalize the dataset;
Step3: Repeat
    {
      Weights = grad (params);
      Update weights;
    } until convergence
Step4: z = dot product of predictor variables and updated weights;
Step5: prediction_limit = sigmoid function (z);
Step6: Predict the target class
```

Artificial Neural Network (ANN):

An Artificial Neural Network (ANN) is a computational model inspired by the way biological neural networks in the human brain process information. It consists of interconnected nodes (neurons) organized in layers. Each neuron receives input, processes it with a specific function (often a weighted sum followed by an activation function), and passes the output to the next layer. ANNs are used for various tasks including pattern recognition, classification, regression, and optimization.

Components of an ANN:

1.Input Layer: Receives the initial data or features.

2.Hidden Layers: Intermediate layers between the input and output layers, where complex computations occur. Each neuron in a hidden layer combines its inputs using weights and applies an activation function to produce an output.

3.Output Layer: Produces the final output based on the processed information from the hidden layers.

Mathematical Formulation:

(Feedforward Process):

Given an input vector x=(x1,x2,…,xn)\mathbf{x} = (x_1, x_2, \ldots, x_n)x=(x1,x2,…,xn), the output yyy of an ANN with LLL layers can be represented as:

For each neuron j **in layer** l**:**

Given an input vector x=(x1,x2,…,xn)\mathbf{x} = (x_1, x_2, \ldots, x_n)x=(x1,x2,…,xn), the output yyy of an ANN with LLL layers can be represented as:

For each neuron j **in layer** l**:**

$$z_j^l = \sum_{k=1}^{N_{l-1}} w_{jk}^l a_k^{l-1} + b_j^l$$

5. Gradient Boosting Algorithm:

Gradient Boosting is a machine learning technique for regression and classification problems, which builds models sequentially by minimizing errors (residuals) of the previous model. It belongs to the ensemble learning methods, where multiple models (often decision trees) are combined to improve predictive performance.

Key Components of Gradient Boosting:

1.Base Learners: Typically decision trees, but can be other models as well.

2.Loss Function: Measures the difference between predicted and actual values.

3.Gradient Descent: Updates the model parameters (weights) to minimize the loss function.

4.Boosting: Sequentially adds models to correct errors made by previous models.

Mathematical Formulation:

Given a training set (xi,yi)(\mathbf{x}_i, y_i)(xi,yi) for i=1,…,Ni = 1, \ldots, Ni=1,…,N, where xi\mathbf{x}_ixi are feature vectors and yiy_iyi are corresponding labels:

Objective: Minimize the loss function L(y,F(x))L(y, F(\mathbf{x}))L(y,F(x)) where F(x)F(\mathbf{x})F(x) is the ensemble model.

Gradient Calculation: Compute the negative gradient of the loss function with respect to the model predictions:

$$F_{\text{new}}(\mathbf{x}) = F_{\text{current}}(\mathbf{x}) + \eta \cdot \text{Tree}(g_i)$$

III.MODULES

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. inka

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.

IV.OUTPUT RESULT

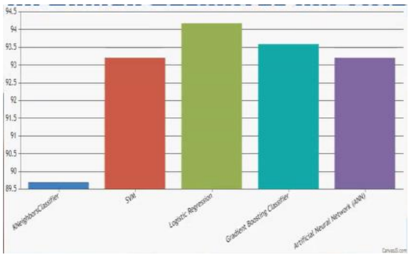


Fig 1 final result graph

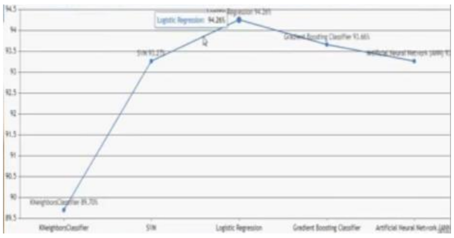


Fig 2 line chart

S.NO	Algorithm	Accuracy
1	K Neighbors Classifier	89.70
2	SVM	93.26
3	Logistic Regression	94.25
4	Gradient Boosting Classifier	93.66
5	ANN	93.26

Fig 3 final accuracy

V.CONCLUSION

In this paper, we present a two-stage job tile identification methodology based on semi-supervised and unsupervised machine learning algorithms with minimal labeling. 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.

VI.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 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.

[16] 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.

[17] 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.

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

[19] F. Colace, M. D. Santo, M. Lombardi, F. Mercorio, M. Mezzanzanica, and F. Pascale, ''Towards labour market intelligence through topic modelling,''in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1–10. [20] E. Mankolli and V. Guliashki, ''A hybrid machine learning method for text analysis to determine job titles similarity,'' in *Proc. 15th Int.Conf. Adv.*

Technol., Syst. Services Telecommun. (TELSIKS), Oct. 2021,pp. 380–385.

[21] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, ''Human resources for Big Data professions: A systematic classification of job roles and required skill sets,'' *Inf. Process. Manage.*, vol. 54, no. 5, pp. 807–817, 2018.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ''Distributed representations of words and phrases and their compositionality,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.

[23] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, ''Advances in pre-training distributed word representations,'' in *Proc.11th Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, May 2018,pp. 1–4.

[24] Y. Bengio, R. Ducharme, and P. Vincent, ''A neural probabilistic language model,'' *J. Mach. Learn. Res.*, vol. 3, pp. 1155–1237, 2003.

[25] R. Collobert and J. Weston, ''A unified architecture for natural language processing: Deep neural networks with multitask learning,'' in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, Jul. 2008, pp. 160–167.

AUTHOR's PROFILE:

[1] Mr.B.Suresh Reddy, currently working as an Assistant Professor in the MCA, at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. His areas of interest are Data structure, machine learning, java, and web technologies.

[2] Ms A.Sirisha currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. He Completed B.Sc. in Computer Science from Dr. Jayapradamba degree college in guntur, Andhra Pradesh. His areas of interests are python & Machine learning.