IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Job Title Identification With BERT And Unsupervised Learning

Y.Harshavardhan[1] Mr.P Ramesh[2]

[1]PG student, Vemu Institute of Technology, P. Kothakota.

[2]Assistant Professor,Vemu Institute of Technology, P.kothakota.

ABSTRACT

This project aims to enhance job title identification in online job advertisements using advanced data science techniques. Current methods heavily rely on labeled datasets focused on the US job market. Our two-stage approach first employs BERT for sector classification, such as Information Technology, followed by unsupervised machine learning to identify the most relevant job title within the predicted sector. We introduce a novel document embedding strategy to address challenges in job ad processing and classification. Experimental results show a significant 14% improvement in accuracy and a 23.5% boost with our document embedding approach, outperforming traditional machine learning methods like SVM, Naïve Bayes, and Logistic Regression. Additionally, we extend the study by experimenting with the CNN2D algorithm, showcasing its ability to achieve higher accuracy by filtering features at multiple neuron iterations.

Keywords: Advertisements, BERT

INTRODUCTION:

Digital transformation and rise of online job portals have generated vast amounts of data that require efficient processing and analysis for valuable insights. Data science emerges as a potent tool for classifying diverse data types, including text from job advertisements, which traditionally posed challenges due to their non-structured nature and diverse lexicon. Existing methods often rely on large labeled datasets tailored to specific markets and struggle with noisy and generic information in job ads. This study introduces a novel approach to job title identification using self-supervised and unsupervised machine learning algorithms, minimizing labeling needs while maintaining high accuracy. Our methodology involves sector classification of job ads using various text classifiers like SVM, Naïve Bayes, Logistic

Regression, and BERT, followed by matching with occupations within the predicted sector. We employ advanced techniques for text vector representation, customized document embedding, and feature selection to enhance classification accuracy. By addressing the limitations of existing methods and focusing on minimal labeling, our approach offers a scalable and adaptable solution for job title identification across different languages and markets.

LITERATURE SURVEY:

F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao *et al*

Here the author introduces Carotene, a machine learning-driven semi-supervised system designed for job title classification in online recruitment. Unlike traditional methods, Carotene employs a diverse range of classification and clustering techniques within a cascade classifier architecture, enhancing scalability for a vast taxonomy of job categories. The system features a two-stage classifier cascade, offering both coarse and fine-level classification. The paper compares Carotene's performance with an earlier flat classifier-based version and contrasts it with a third-party occupation classification system. Additionally, experimental results from real-world industrial data, evaluated through machine learning metrics and user experience surveys, validate Carotene's effectiveness and scalability in accurately classifying job titles for optimal job-seeker matching.

I. Rahhal, K. Carley, K. Ismail, and N. Sbihi*et al*

As per the author proposes a solution to address the disconnect between university curricula and current job market demands, particularly focusing on the IT sector. Recognizing that many students lack access to up-to-date information about job market needs, this study utilizes data from job portals and university websites. Machine learning algorithms classify job ads by occupation, while text analysis extracts required skills and qualifications. The research identifies programming as a high-demand occupation in IT, with emerging roles like data scientists seeing a significant increase in job openings. A comparative analysis reveals slight mismatches between university offerings and job market demands. The study concludes with the development of a Dashboard to guide students towards career paths with better employment prospects.

F. Amato, R. Boselli, M. Cesarini, F. Mercorio*et al*

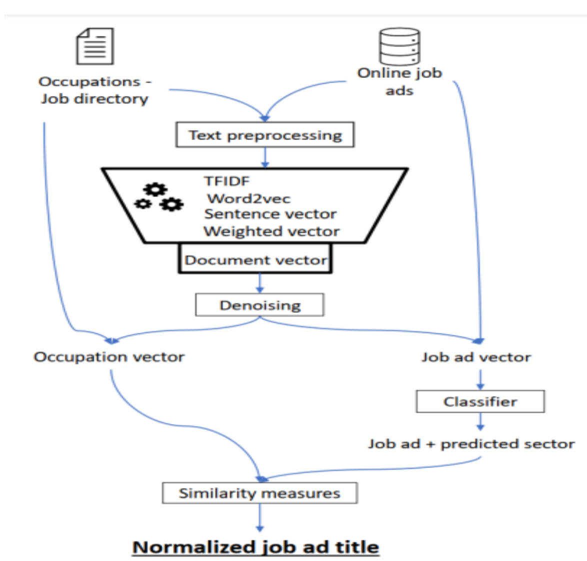
Since the author proposes a solution to address the disconnect between university curricula and current job market demands, particularly focusing on the IT sector. Recognizing that many students lack access to up-to-date information about job market needs, this study utilizes data from job portals and university websites. Machine learning algorithms classify job ads by occupation, while text analysis extracts required skills and qualifications. The research identifies programming as a high-demand occupation in IT, with emerging roles like data scientists seeing a significant increase in job openings. A comparative analysis reveals slight mismatches between university offerings and job market demands. The study concludes with the development of a Dashboard to guide students towards career paths with better employment prospects.

PROBLEM STATEMENT:

Data science algorithms often used to extract useful knowledge from unstructured text data such as Identifying Job Title by analysing Job Text Description. All existing algorithms are heavily dependent on large Label data for perfect classification and gathering huge label require lots of experience and time. All existing algorithms were using Occupational Information Network (O*NET) data from USjob market and this existing algorithm were not applying any additional technique to improve accuracy.

PROPOSED METHOD:

ARCHITECTURE:



To overcome above problem author of this paper employing two stage Job Title Identification and this stages include

In first Stage author using Bidirectional Encoder Representations from Transformers (BERT) to first classify thejob 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.

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 amount of labels and doesn't require huge labels of data.

Propose BERT model has compared with many other vector models like TFIDF, WORD2VEC etc. Compare to TFIDF and WORD2VEC propose BERT with Euclidean distance is giving high accuracy.

JOB TITLE DATASET:

To train all algorithms author has generated his own dataset but not publish on internet so we have used Job Title Description dataset from KAGGLE and below are the dataset details

```
ID,Query,Job Title,Description
1,Data Scientist,Junior Data Scientist Apprenticeship,"Job Description As a Junior Data Scientist at IBM, you will work as part of a tean
2,Data Scientist,"HBO Data Scientist, Content Science","OVERALL SUMMARY As a Data Scientist on the Data Science Solutions team, t
3,Data Scientist, Junior Data Scientist,"The Team: The Data science team is a newly formed applied research team within S&P; Globa
4,Data Scientist,Jr Data Scientist,"We now have a need for junior Data Scientist(s) in the NY area (or remote). The successful candida
5,Data Scientist," Data Scientist, Premium Content","Do you want to help guide the core business of Spotify using insights from analys
6,Data Scientist, Data Scientist,"The Mayor's Office of Data Analytics (MODA) is New York City's ideas incubator for operational analytic
7,Data Scientist,Customer Data Scientist,"Smartly.io is a fast-growing company aiming to be the best and nicest place for the most tal
8,Data Scientist,Data Analyst / Data Scientist,We have a client that is looking for a data scientist to add to its team. This position is for
9,Data Scientist,Data Scientist,"Marketing Statement MetroPlus Health Plan provides the highest quality healthcare services to resident
10,Data Scientist, Junior Data Scientist,"Job Description: Junior Data Scientist, Membership Analytics Job Description: We are looking
11,Data Scientist,Data Scientist,"About Narrativ Narrativ [https://narrativ.com/] is a NYC-based, NEA-backed tech startup that is buildi
12,Data Scientist,Geospatial Data Scientist,"NYC Department of Finance (DOF) is responsible for administering the tax revenue laws of
13,Data Scientist,Associate Data Scientist,"Overview: Alliant is seeking an Associate Data Scientist to join its Data Science team to build
14,Data Scientist,Data Scientist,"_Position Mission/Summary:_ Looking for opportunities to use cutting edge technologies analyzing pe
15,Data Scientist, Data Scientist,"Overview: Alliant is seeking a Data Scientist to join its Data Science team to build predictive models a
16,Data Scientist,Data Scientist,"Join Sweeten, a fast growing, award-winning tech company based in New York City and founded by
17,Data Scientist,"Data Scientist, Patient Messaging","\--------- The Role: \--------- Phreesia's patient messaging platform delivers targe
18,Data Scientist, Distinguished Data Scientist,"As a Distinguished Data Scientist, you will provide data science and machine learning l
19,Data Scientist, Data Scientist,"Overview Data & Analysis – Data Scientist Digitas is a highly-caffeinated playground where brilliant m
20,Data Scientist, Data Scientist/ Researcher,"As the Data Scientist with the Market Intelligence team, you will support a team of resea
21,Data Scientist,Data Scientist,"Data science is the core around which Custora is built. The central value proposition of our product is
22,Data Scientist,Data Warehouse Intern,"Newsela is an Instructional Content Platform that brings together engaging, accessible cont
23,Data Scientist, Data Scientist,"FreeWheel, a Comcast company, has superior end-to-end technology, premium marketplace, and be
24,Data Scientist, Data Scientist,"Vroom.com is a venture-backed, fast-growing start-up focused on revolutionizing the car buying exp
```

In above dataset first row contains dataset column names and remaining rows contains dataset values and in dataset we can see Job Title, Name and Description and by using above dataset we will train and test all algorithm performance.

METHODOLOGY:

Python Classes and Packages Import

The project kicks off by importing pivotal Python libraries and classes. These include packages for data manipulation, natural language processing (NLP), and the implementation of machine learning models.

Text Preprocessing

To ensure the accuracy of job title predictions, the job descriptions are preprocessed. This involves removing stop words, special characters, and other irrelevant elements. By doing so, the text data is streamlined and becomes more suitable for analysis.

Dataset Exploration

The dataset, consisting of job descriptions, is loaded and visualized to understand its structure and content. Through exploratory data analysis, the distribution of job titles is examined, and potential patterns or trends within the data are identified.

Graph Plotting for Job Titles

A graphical representation is crafted to display the frequency distribution of various job titles within the dataset. This visualization provides insights into the prominence of different job roles, with job titles plotted on the x-axis and their respective counts on the y-axis.

Feature Extraction using BERT and TFIDF

Feature extraction is a critical step in converting textual data into a format suitable for machine learning. Here, both BERT (Bidirectional Encoder Representations from Transformers) and TFIDF (Term Frequency-Inverse Document Frequency) methods are employed to transform the job descriptions into numeric vectors, which are then utilized for training machine learning models.

Normalization and CHI2 Algorithm

Post feature extraction, the features are normalized to ensure uniformity and comparability across different scales. Subsequently, the CHI2 algorithm is implemented to enhance the relevance and significance of the features, thereby boosting the predictive power of the machine learning models.

Data Splitting and Model Evaluation

The dataset is partitioned into training and testing subsets to train the machine learning models and validate their performance. Evaluation metrics, including accuracy, precision, recall, and confusion matrices, are computed to gauge the efficacy of each model.

Model Training and Evaluation

Various machine learning algorithms, encompassing SVM, Naïve Bayes, Logistic Regression, BERT, and the CNN2D extension, are trained and assessed using the features extracted earlier. The obtained accuracy and other performance metrics are scrutinized to pinpoint the most proficient method for predicting job titles.

Performance Visualization

The performance of each algorithm is graphically depicted, with algorithm names plotted on the x-axis and accuracy and other metrics on the y-axis. Additionally, a tabular representation is employed to juxtapose the performance of each algorithm for comparative analysis.

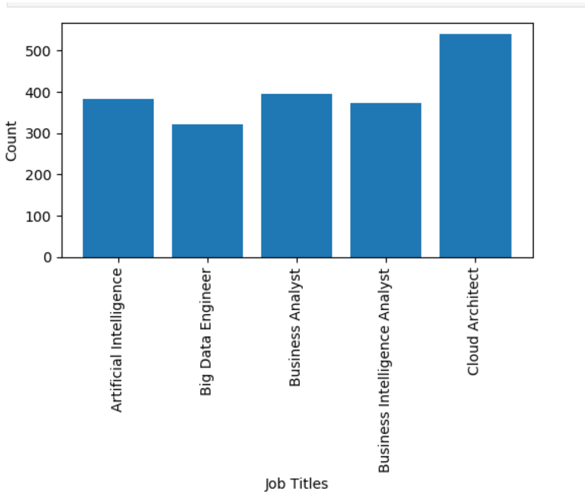
Prediction on Test Data

In the concluding phase, the trained models are deployed to forecast job titles based on the test dataset's job descriptions. The predicted job titles are juxtaposed with the actual titles to determine the models' real-world applicability and accuracy.

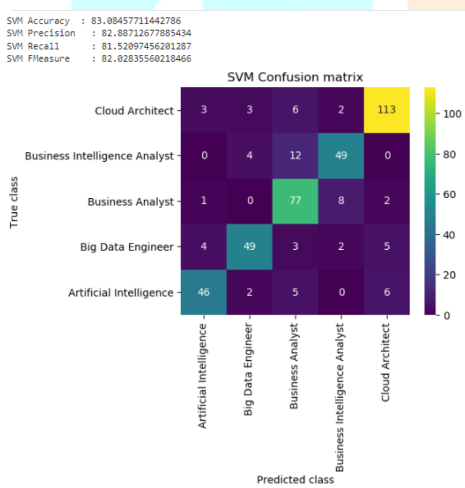
Extension: Exploring Advanced Algorithms

While the original paper primarily focused on conventional machine learning algorithms like SVM, Naïve Bayes, and Logistic Regression, we extended the research by incorporating advanced algorithms like CNN2D. This convolutional neural network-based algorithm refines features through multiple neuron iterations, ensuring the model is trained with optimal features. This enhancement has demonstrated a significant improvement in accuracy, showcasing the potential of advanced algorithms in optimizing job title prediction systems.

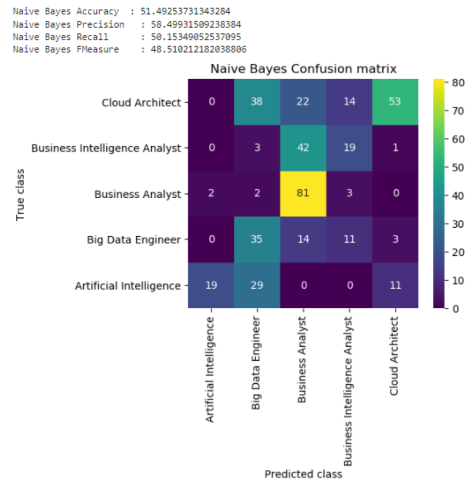
RESULTS:



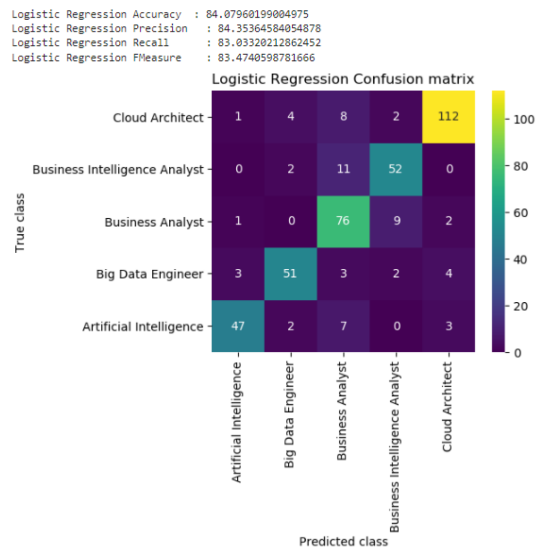
Finding and plotting graph of various JOBS found in dataset where x-axis represents JOB TITLE and y-axis represents counts



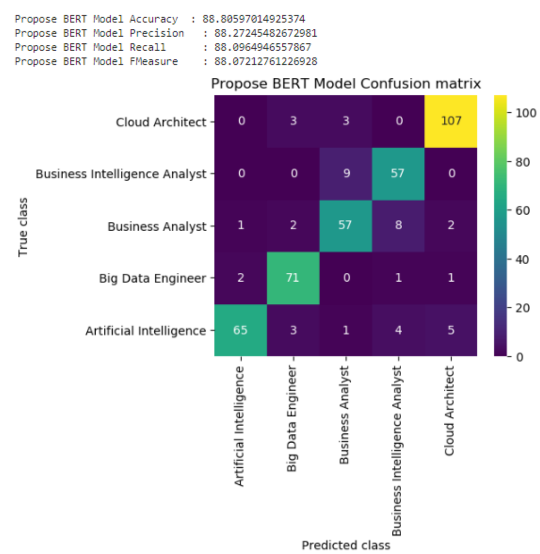
Training SVM on TFIDF features and it got 83% accuracy and can see other metrics also and in confusion matrix graph x-axis represents True Job Title and Y-axis represents Predicted Job Title and all different colour boxes in diagnol represents correct prediction count and remaining blue boxes contains incorrect prediction count



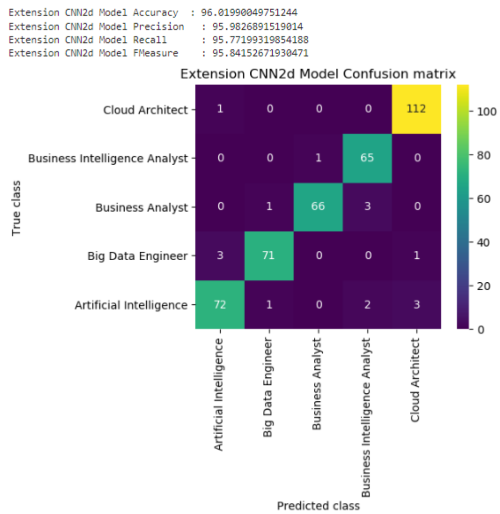
In above Naïve Bayes got 51% accuracy



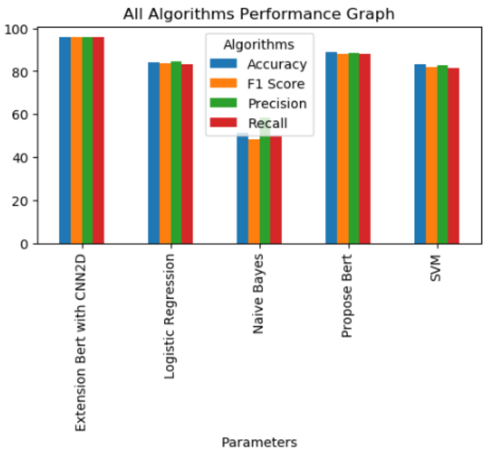
In above Logistic Regression got 84% accuracy



In above screen propose BERT model with max similarity measure got 88% accuracy which is higher than existing algorithms and can see other metrics also



In above screen extension CNN2D model got 96% accuracy and can see other metrics also



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

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	SVM	82.887127	81.520975	82.028356	83.084577
1	Naive Bayes	58.499315	50.153491	48.510212	51.492537
2	Logistic Regression	84.353646	83.033202	83.474060	84.079602
3	Propose BERT	88.272455	88.096495	88.072128	88.805970
4	Extension BERT CNN2D	95.982689	95.771993	95.841527	96.019900

In above screen displaying all algorithms performance in tabular format

Prediction:

```
Job Description = Note: By applying to this position your application is automatically submitted to the following locations: Los Angeles, CA, USA; Ann Arbor, MI, USA; C
PREDICTED JOB TITLE =====> Cloud Architect

Job Description = Overview Belstone is a fast-growing, technology-led merchant bank that drives capital to the private companies fueling economies, creating new jobs, a
PREDICTED JOB TITLE =====> Artificial Intelligence

Job Description = Marketing used to be an exercise in one-to-many communication: billboards, magazine ads, and - more recently - having a powerful social media presence
PREDICTED JOB TITLE =====> Big Data Engineer

Job Description = Job Title: Industrial Engineer - Data Operations Intern Reports To: Manager, Data Operations Department: Data Operations FLSA Status: Non-exempt Hours
PREDICTED JOB TITLE =====> Big Data Engineer

Job Description = OVERALL SUMMARY: Reports to the Senior Manager of Business Continuity and Crisis Management (BCCM). The BCCM Team helps to prepare employees for unpla
PREDICTED JOB TITLE =====> Cloud Architect

Job Description = Since 1851, MassMutual?s commitment has always been to help people protect their families, support their communities, and help one another. This is wh
PREDICTED JOB TITLE =====> Cloud Architect

Job Description = Job Description Corporate IT is searching for a Business Analyst within our End User Technology (EUT) team. As a key member of the team, you will prov
PREDICTED JOB TITLE =====> Business Analyst

Job Description = Essential duties and responsibilities include the following. Other duties may be assigned. Interprets historical, current, and projected data to ident
PREDICTED JOB TITLE =====> Big Data Engineer

Job Description = OVERALL SUMMARY: HBO's Digital Products division is responsible for the entire technology stack that enables our customers to access HBO's programming
PREDICTED JOB TITLE =====> Cloud Architect

Job Description = Architect- Cloud Location: Virtual (very minimal travel). Unisys is a global information technology company that builds high performance, security-cen
PREDICTED JOB TITLE =====> Cloud Architect
```

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 project systematically employed Python libraries for text preprocessing, dataset exploration, and model training. It began with importing necessary packages and defining code to clean text data. Exploratory data analysis included displaying job dataset values and plotting graphs to visualize job title distribution. BERT and TFIDF vectors were created from job descriptions, normalized, and subjected to the CHI2 algorithm. Various models including SVM, Naïve Bayes, Logistic Regression, and proposed BERT model were trained and evaluated. The extension CNN2D model achieved high accuracy. Performance metrics were displayed graphically and in tabular format. Test data predictions demonstrated effective job title prediction capabilities.

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. ACM SIGIR Conf. Res. Develop. Inf. Retr., Aug. 2015, pp. 1101–1104.

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

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

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

[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, ‘‘Towardslabour 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