Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

A Comparative Study of BERT and CNN2D Techniques for Job Title Identification

[1]E.RADHAKRISHNA, [2] SATISH REDDY KANALA
[1]M.Tech student, [2] Associate Professor
[1,2]Department of CSE
[1,2] Sri Sai Institute Of Technology And Science , Masapeta, Rayachoty (M), Annamayya Dt ,AP.
ssitsrk@gmail.com, kanalasatishreddy5@gmail.com

ABSTRACT

Data science techniques are increasingly used to extract insights from large datasets, particularly in analyzing job market trends by classifying online job advertisements. Traditional multi-label classification methods, like self-supervised learning and clustering, have shown promise but often require extensive labeled datasets and focus on specific databases such as O*NET, which is tailored to the US job market. This paper introduces a two-stage job title identification methodology designed for smaller datasets. It utilizes Bidirectional Encoder Representations from Transformers (BERT) to classify job ads by sector and then applies unsupervised learning and similarity measures to match job titles within the predicted sector. The proposed document embedding strategy, incorporating weighting and noise removal, enhances accuracy by 23.5% compared to Bag of Words models. Results indicate a 14% improvement in job title identification accuracy, achieving over 85% in certain sectors. The study also explores the use of CNN2D, an advanced algorithm, to further enhance classification performance by filtering features through multiple neural network iterations.

Index Terms: job market, BERT

INTRODUCTION:

The rapid expansion of the Internet and the rise of social media have led to an enormous amount of data, demanding swift and efficient processing to extract valuable insights for decision-making. Data science techniques play a crucial role in this context by enabling the analysis and classification of diverse data types, such as text, images, and video, and can significantly improve upon traditional, resource-intensive methods.

The job market has similarly transitioned to online platforms, with employers and recruiters posting job advertisements across various websites to reach a broader audience. This digital shift presents a valuable opportunity to analyze job market trends and understand the specific needs in terms of skills and occupations. Such insights are beneficial not only for labor market analysts and policymakers aiming to enhance employment strategies but also for job seekers and students seeking relevant career opportunities and necessary training.

LITERATURE SURVEY:

G. Mezzour *et al*

Offshore sector in Morocco offers numerous job opportunities, but analyzing related job ads is challenging due to their unstructured nature. This study examines job ads from February to August 2017, utilizing machine learning and text mining techniques. We analyze required skills, including natural and programming languages, education level, experience, contract types, and salaries. Our findings highlight that French is crucial for offshore roles, with English and Spanish also valued. Development and web design are key IT roles, with Java, SQL, JavaScript, and PHP being the most sought-after programming languages.

V. Guliashki *et al*

This paper presents a hybrid approach that merges k-NN and SVM machine learning techniques to identify job titles with similar descriptions and industries. This innovative method enhances both the accuracy and efficiency of the candidate selection process, streamlining the task of matching job titles to suitable candidates. By integrating these two methods, the approach improves the overall effectiveness of job title classification, making it faster and more precise.

PROBLEM STATEMENT:

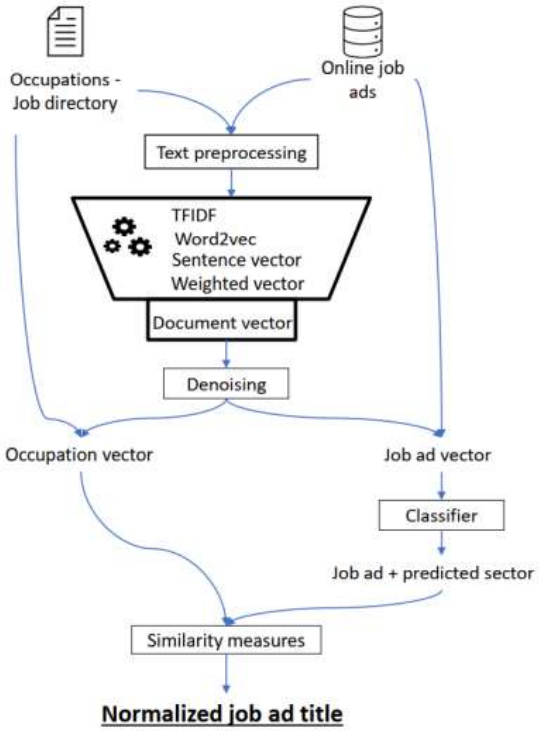
Data science algorithms often used to extract useful knowledge from unstructured text data such as Identifying Job Title by analysing Job Text Description. All existing algorithms are heavily dependent on large Label data for perfect classification and gathering huge label require lots of experience 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:

This paper introduces a two-stage method to tackle the complexities of job title identification. Initially, Bidirectional Encoder Representations from Transformers (BERT) are utilized to classify job advertisements into specific sectors, such as Information Technology or Agriculture. BERT translates unstructured text into numerical vectors while preserving semantic meaning. In the subsequent stage, the Euclidean Distance algorithm measures the similarity between job ads and potential job titles, identifying the closest match, even

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

with limited labels. This BERT-based approach, paired with Euclidean Distance, surpasses traditional models like TFIDF and WORD2VEC, offering enhanced accuracy in job title identification.



ARCHITECTURE

TIME SERIES 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:

Text Preprocessing

Text preprocessing is a critical step in preparing job descriptions for analysis and model training. The goal is to clean the text data by removing elements that do not contribute to the prediction of job titles. This involves:

- **Removing Stop Words**: Words such as "the," "and," or "is" are common but do not provide meaningful information for the prediction task. Removing them helps in focusing on more relevant terms.
- **Eliminating Special Symbols**: Symbols like punctuation marks or special characters that do not contribute to the text's semantic meaning are discarded.
- **Stripping Irrelevant Elements**: Any other irrelevant elements, such as extra spaces or HTML tags, are removed to ensure the text data is clean and formatted consistently.

This preprocessing ensures that the text data is in a suitable form for subsequent analysis and feature extraction.

Dataset Exploration

Exploring the dataset is crucial for understanding its structure and content. This involves:

- **Reading the Dataset**: The job descriptions dataset is loaded into a DataFrame, allowing for examination of its structure, including columns, data types, and sample entries.
- **Understanding Distribution**: Initial analysis helps in understanding how job titles are distributed across the dataset. This might include examining the frequency of different job titles and identifying any imbalances or biases in the data.

Exploratory analysis provides insights into the data’s characteristics and helps in tailoring the preprocessing and feature extraction steps to improve model performance.

Graph Plotting for Job Titles

Visualizing the distribution of job titles helps in understanding their frequency and prevalence within the dataset. This is achieved through:

- **Graph Plotting**: A bar graph is plotted, where the x-axis represents various job titles and the y-axis shows their respective counts. This visualization offers a clear view of the most common and rare job titles, highlighting any trends or anomalies.

Such graphs are instrumental in identifying which job titles are most prevalent and whether any job titles are underrepresented, which might impact the model's training.

Feature Extraction using BERT and TF-IDF

Feature extraction transforms text data into a numeric format that machine learning models can process:

- **BERT (Bidirectional Encoder Representations from Transformers)**: BERT is a sophisticated NLP model that captures context from both directions in a sentence, providing rich, contextualized word embeddings. It generates high-dimensional vectors for job descriptions, capturing nuanced meanings and relationships in the text.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**: TF-IDF is a statistical measure that reflects the importance of a word in a document relative to a collection of documents. It transforms text into numeric vectors based on word frequency and document rarity, helping in identifying significant terms.

Both BERT and TF-IDF are applied to the job descriptions to convert them into feature vectors suitable for machine learning model training.

Normalization and CHI2 Algorithm

Normalization: After feature extraction, the next step is to normalize the features to ensure that all variables contribute equally to the model's performance. This step adjusts the scale of features, making them comparable and improving the stability and convergence of machine learning models.

CHI2 Algorithm: The CHI2 (Chi-squared) test is applied to the features to evaluate their importance. This statistical test assesses the independence of features from the target variable, helping to select the most relevant features and improve model accuracy by focusing on the most significant ones.

Data Splitting and Model Evaluation

Data Splitting: The dataset is divided into training and testing sets. Typically, an 80-20 split is used, where 80% of the data is used for training the model, and 20% is reserved for testing. This ensures that the model is evaluated on unseen data, providing a realistic measure of its performance.

Model Evaluation: Various evaluation metrics are computed to assess the model's effectiveness:

- **Accuracy**: Measures the proportion of correctly predicted job titles.
- **Precision**: Indicates how many of the predicted job titles were correct.
- **Recall**: Reflects how many of the actual job titles were correctly identified.
- **Confusion Matrices**: Provide a visual representation of prediction results, showing true positives, false positives, true negatives, and false negatives.

These metrics help in understanding the model's performance and identifying areas for improvement.

Model Training and Evaluation

Different machine learning algorithms are trained and evaluated using the extracted features:

- **SVM (Support Vector Machine)**: A powerful classifier that finds the optimal hyperplane to separate different job titles.
- **Naïve Bayes**: A probabilistic classifier based on Bayes' theorem, effective for text classification tasks.
- **Logistic Regression**: A statistical model used for binary classification, which can be extended to handle multi-class problems.
- **BERT**: Fine-tuned for the job title prediction task, leveraging its contextual understanding of text.
- **CNN2D (Convolutional Neural Network)**: Although typically used for image analysis, CNNs can be adapted for text classification by treating text as a sequence of data.

The performance of each model is analyzed using metrics like accuracy, precision, recall, and confusion matrices to determine the most effective approach for job title prediction.

Performance Visualization

The performance of various algorithms is visualized to facilitate comparison:

- **Graphs**: Bar graphs or line plots display the accuracy and other metrics of different models. The x-axis represents the names of the algorithms, while the y-axis shows their performance metrics.
- **Tabular Format**: A table is used to present the performance metrics of each model, allowing for easy comparison and evaluation.

These visualizations help in understanding which algorithms perform best and in making data-driven decisions about model selection.

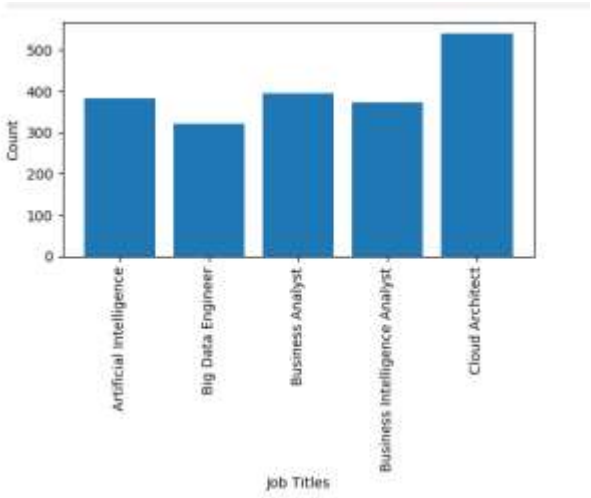
Prediction on Test Data

The final step involves using the trained models to predict job titles based on job descriptions from the test dataset. The predicted titles are compared with the actual titles to evaluate the model's real-world performance and effectiveness in accurately classifying job descriptions.

Extension

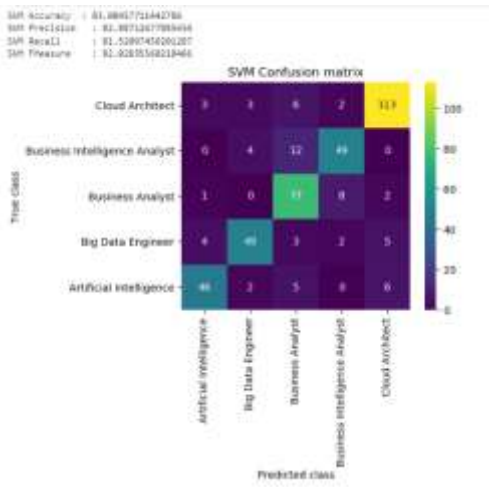
In the proposed paper, traditional machine learning algorithms like SVM, Naïve Bayes, and Logistic Regression were employed, but advanced algorithms like CNN2D and Bi-LSTM were not explored. As an extension, CNN2D has been experimented with in this work. CNN2D filters features through multiple neuron iterations, allowing the model to train with the most relevant features, which helps in achieving higher accuracy. This exploration of advanced algorithms demonstrates their potential in improving job title prediction and offers valuable insights into their effectiveness compared to traditional methods.

RESULTS:

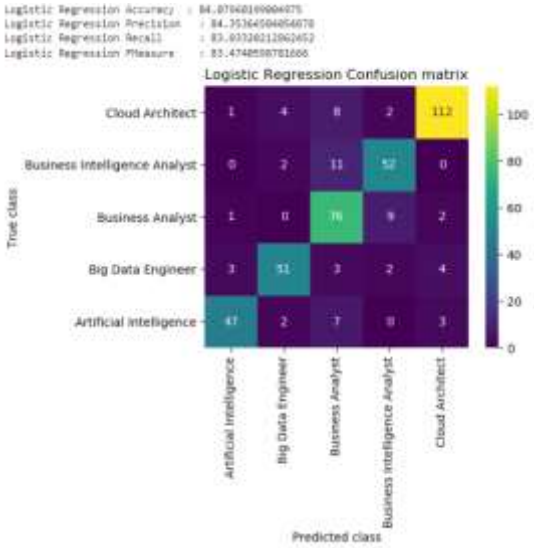


Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

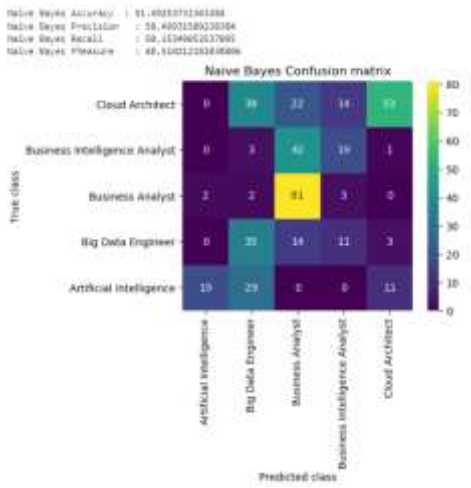
Finding and plotting graph of various JOBS found in dataset



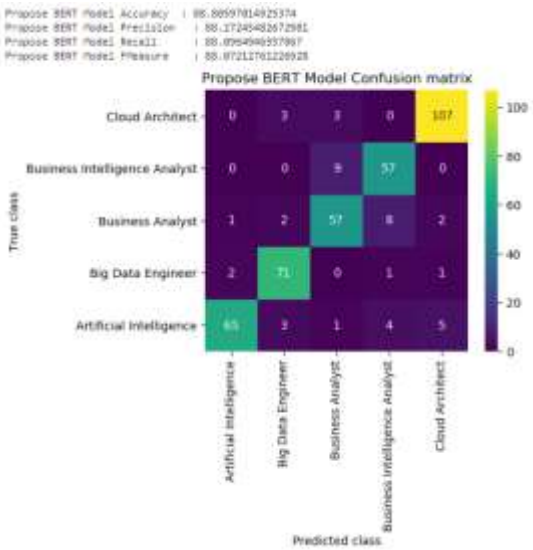
Training SVM on TFIDF features and it got 83% accuracy



Training Logistic Regression got 84% accuracy

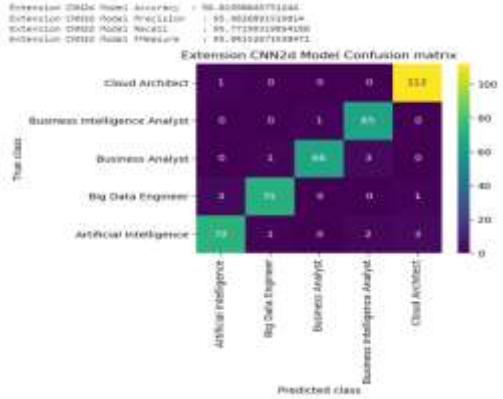


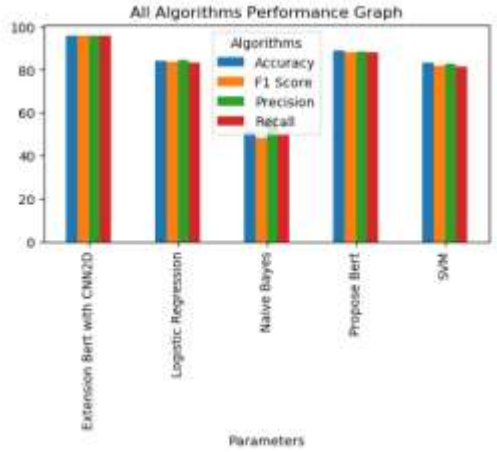
Training Naïve Bayes got 51% accuracy



Training BERT model with max similarity measure got 88% accuracy

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024





Training CNN2D model got 96% accuracy

x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars

Prediction:

Predicted JOB title as Big data Engineer or Cloud Architect

CONCLUSION

This project utilized Python libraries for systematic text preprocessing, dataset exploration, and model training. Starting with the import of necessary packages, text data was cleaned and prepared. Exploratory data analysis included displaying job dataset values and plotting graphs to visualize job title distribution. Job descriptions were transformed into BERT and TFIDF vectors, normalized, and analyzed using the CHI2 algorithm. Models such as SVM, Naïve Bayes, Logistic Regression, and the proposed BERT model were trained and evaluated. The CNN2D model exhibited high accuracy. Performance metrics were presented graphically and in tables, demonstrating effective job title prediction on test data.

REFERENCES:

[1] C. P. Veiga, ‘‘Analysis of quantitative methods of demand forecasting: Comparative study and financial performance,’’ Ph.D. dissertation, Dept. Manag., Pontifical Catholic Univ. Paraná, Curitiba, Brazil, 2009.
[2] J. H. Friedman, ‘‘Greedy function approximation: A gradient boosting machine,’’ Ann. Statist., vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T.-Y. Liu, ‘‘LightGBM: A highly efficient gradient boosting decision tree,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 3146–3154.
[4] T. Chen and C. Guestrin, ‘‘XGBoost: A scalable tree boosting system,’’ in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Aug. 2016, pp. 785–794.
[5] A. V. Dorogush, V. Ershov, and A. Gulin, ‘‘CatBoost: Gradient boosting with categorical features support,’’ 2018, arXiv:1810.11363.
[6] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[7] S. Yadav, T.-M. Choi, S. Luthra, A. Kumar, and D. Garg, ‘‘Using Internet of Things (IoT) in agri-food supply chains: A research framework for social good with network clustering analysis,’’ IEEE Trans. Eng. Manag., vol. 70, no. 3, pp. 1215–1224, Mar. 2023, doi: 10.1109/TEM.2022.3177188.
[8] J. Zheng, L. Wang, L. Wang, S. Wang, J.-F. Chen, and X. Wang, ‘‘Solving stochastic online food delivery problem via iterated greedy algorithm with decomposition-based strategy,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 2, pp. 957–969, Feb. 2023, doi: 10.1109/TSMC. 2022.3189771.
[9] V. K. Shrivastava, A. Shrivastava, N. Sharma, S. N. Mohanty, and C. R. Pattanaik, ‘‘Deep learning model for temperature prediction: A case study in New Delhi,’’ J. Forecasting, vol. 43, no. 1, Feb. 2023, doi: 10.1002/for.2966.
[10] Y. Zhang, L. Wang, X. Chen, Y. Liu, S. Wang, and L. Wang, ‘‘Prediction of winter wheat yield at county level in China using ensemble learning,’’ Prog. Phys. Geogr., Earth Environ., vol. 46, no. 5, pp. 676–696, Oct. 2022.
[11] I. Shah, F. Jan, and S. Ali, ‘‘Functional data approach for short-term electricity demand forecasting,’’ Math. Problems Eng., vol. 2022, Jun. 2022, Art.no. 6709779.
[12] F. Lisi and I. Shah, ‘‘Forecasting next-day electricity demand and prices based on functional models,’’ Energy Syst., vol. 11, no. 4, pp. 947–979, Nov. 2020.
[13] I. Shah, H. Iftikhar, and S. Ali, ‘‘Modeling and forecasting electricity demand and prices: A comparison of alternative approaches,’’ J. Math., vol. 2022, Jul. 2022, Art. no. 3581037.
[14] I. Shah, S. Akbar, T. Saba, S. Ali, and A. Rehman, ‘‘Short-term forecasting for the electricity spot prices with extreme values treatment,’’ IEEE Access, vol. 9, pp. 105451–105462, 2021.
[15] I. Shah, H. Bibi, S. Ali, L. Wang, and Z. Yue, ‘‘Forecasting one-dayahead electricity prices for Italian electricity market using parametric and nonparametric approaches,’’ IEEE Access, vol. 8, pp. 123104–123113, 2020.
[16] N. Bibi, I. Shah, A. Alsubie, S. Ali, and S. A. Lone, ‘‘Electricity spot prices forecasting based on ensemble learning,’’ IEEE Access, vol. 9, pp. 150984–150992, 2021.
[17] E. S. Gardner Jr., ‘‘Exponential smoothing: The state of the art,’’ J. Forecasting, vol. 4, no. 1, pp. 1–28, 1985.
[18] C. C. Holt, ‘‘Forecasting seasonals and trends by exponentially weighted moving averages,’’ Int. J. Forecasting, vol. 20, no. 1, pp. 5–10, 2004.
[19] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control. San Francisco, CA, USA: Holden Day, 1970.
[20] P. Ramos, N. Santos, and R. Rebelo, ‘‘Performance of state space and ARIMA models for consumer retail sales forecasting,’’ Robot.Comput.-Integr. Manuf., vol. 34, pp. 151–163, Aug. 2015.