Enhancing Job Title Identification With BERT And Unsupervised Learning

Y.Harshavardhan1 Mr.P Ramesh2
1PG student, Vemu Institute of Technology, P. Kothakota.
2Assistant 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 US job market and this existing algorithm were not applying any additional technique to improve accuracy.

PROPOSED METHOD:

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 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.

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.

ARCHITECTURE:
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

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 diagonal 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

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Precision</th>
<th>Recall</th>
<th>FScore</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>82.887127</td>
<td>81.520975</td>
<td>82.028356</td>
<td>83.084577</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>58.499315</td>
<td>50.153491</td>
<td>48.510212</td>
<td>51.492537</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>84.353646</td>
<td>83.033202</td>
<td>83.474060</td>
<td>84.079602</td>
</tr>
<tr>
<td>Propose BERT</td>
<td>88.272455</td>
<td>88.096495</td>
<td>88.072128</td>
<td>88.805970</td>
</tr>
<tr>
<td>Extension BERT CNN2D</td>
<td>95.982689</td>
<td>95.771993</td>
<td>95.841527</td>
<td>96.019900</td>
</tr>
</tbody>
</table>

In above screen displaying all algorithms performance in tabular format
Prediction:

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:


