Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2020, 7(3):300-307**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Efficient Job Title Classification Using FastText and Multi-Stage Classification Pipeline

Saandeep Sreerambatla

Abstract This study develops an efficient job title classification system using a large-scale database of job titles and descriptions. From an initial dataset of 36 million records, we extracted 2.8 million unique job titles and descriptions. We utilized FastText to create a classification algorithm for 2500 predefined classes. A multi-stage classification pipeline was implemented, starting with a classifier for the majority classes and an "unknown" class. Subsequent stages refined the classification for the remaining titles. The labeled dataset was created using word2vec similarity measurements and manual classification. Our models achieved a per-class recall greater than 0.8 for about 1800 classes. This system enhances job matching, career planning, and workforce analysis, providing significant benefits for job seekers, employers, and workforce development professionals. The research demonstrates the potential of combining machine learning techniques with hierarchical classification to tackle large-scale, multi-class classification challenges, highlighting the importance of standardized job descriptions and integrated databases in supporting labor market intelligence and policy-making.

Keywords Job Title Classification, FastText, Multi-Stage Classification Pipeline

1. Introduction

In today's rapidly evolving labor market, the accurate classification of job titles is crucial for efficient human resource management and strategic workforce planning. The proliferation of diverse job titles and descriptions across various industries poses significant challenges to achieving standardized classification. Accurate job classification is essential for numerous organizational processes, including streamlining recruitment and hiring processes, ensuring effective salary benchmarking, making informed promotion decisions, and overall performance management.

Accurate job classification enhances recruitment efficiency by matching job postings with the required skills and qualifications, thereby reducing time-to-hire and minimizing the effort spent on candidate screening. It also supports salary benchmarking by enabling organizations to compare roles within the industry, ensuring competitive and equitable compensation packages. Furthermore, clear and accurate job descriptions aid in performance management by defining roles, responsibilities, and expectations, facilitating fair evaluations and informed promotion decisions. Strategic workforce planning benefits from standardized job descriptions as they help identify skill gaps, design targeted training programs, and develop succession plans.

To address these challenges, we developed a robust job title classification system using a large-scale dataset and advanced machine learning techniques. Starting with an initial dataset of 36 million job records, we extracted 2.8 million unique job titles and descriptions. Utilizing FastText, a library designed for efficient text classification and representation learning, we created a classification algorithm capable of categorizing these titles into 2500 predefined classes.

Our approach involved an innovative multi-stage classification pipeline. Initially, we trained a classifier to identify the majority classes and an "unknown" class, effectively filtering out the most frequent classes.

Subsequent classification stages were employed to refine the categorization of the remaining titles. This hierarchical method enabled us to manage class imbalance and enhance the precision and recall of less frequent classes.

By combining machine learning techniques with hierarchical classification, our models achieved significant performance improvements. Specifically, we achieved a per-class recall greater than 0.8 for approximately 1800 classes. The developed classification system provides valuable insights for job seekers, employers, and workforce development professionals, facilitating better job matching, career planning, and labor market analysis. Additionally, this system helps in maintaining a standardized approach to job classification, supporting more accurate and effective human resource management and strategic planning. Rest of the paper is organised as follows:

- Section 2 presents the background, emphasizing the importance of job classification for industries. It discusses how accurate classification impacts hiring efficiency, salary benchmarking, and promotion decisions, which are essential for aligning workforce capabilities with organizational goals.
- Section 3 reviews related work, exploring various approaches and advancements in job classification using machine learning and NLP techniques. It covers traditional methods like ONET and ESCO, as well as recent innovations in text classification, highlighting the strengths and limitations of these approaches.
- Section 4 details our approach, including the steps taken in data preparation, model building, and the implementation of the multi-stage classification pipeline. This section explains how we handled the large dataset, ensured high-quality training data, and developed a robust classification algorithm using FastText.
- Section 5 discusses our results, providing a thorough analysis and comparison with previous methods. It highlights the performance metrics of our models, such as precision and recall, and examines reasons for the model's performance on certain challenging classes, particularly those related to less frequent job titles.
- Section 6 concludes the paper, summarizing the key findings and their implications for job classification. It also proposes future work to further enhance the accuracy and application of job classification systems, such as incorporating additional data sources and exploring new techniques to address current limitations.

2. Background

The section includes the background of why we need to do the job classification for multiple benefits for organizations.

2.1 Importance of Job Classification

Accurate job classification is crucial for efficient human resource management and strategic workforce planning. The diversity and inconsistency of job titles across industries create significant challenges in achieving standardized classification. Effective job classification systems are essential for various organizational processes, including recruitment, salary benchmarking, performance management, and workforce planning. As depicted in Figure 1, many organizations struggle with inconsistent job titles, leading to numerous problems. Implementing consistent classification and standardization practices is necessary to address these issues.

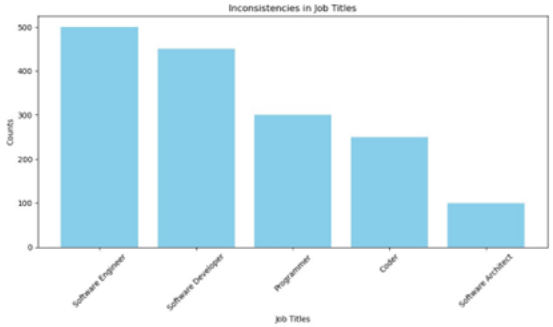

Figure 1: No of Job titles in an organization representing similar

2.2 Impact on Hiring Efficiency

The recruitment process is significantly enhanced by accurate job classification. Matching job postings with the required skills and qualifications becomes more streamlined, reducing time-to-hire and minimizing the effort spent on candidate screening. This efficiency not only helps organizations fill positions more quickly but also ensures that candidates with the right skill sets are identified and hired, thereby improving overall productivity.

2.3 Salary Benchmarking

Accurate job classification supports salary benchmarking by enabling organizations to compare roles within the industry, ensuring competitive and equitable compensation packages. Standardized job titles and descriptions allow for a clearer comparison of roles and responsibilities, helping organizations to offer salaries that attract and retain top talent while maintaining fairness and consistency.

2.4 Performance Management

Clear and accurate job descriptions are critical for effective performance management. They define roles, responsibilities, and expectations, facilitating fair evaluations and informed promotion decisions. With standardized job classifications, managers can set clear performance metrics and goals, ensuring that employees understand their responsibilities and how their performance will be measured.

2.5 Strategic Workforce Planning

Strategic workforce planning benefits from standardized job descriptions as they help identify skill gaps, design targeted training programs, and develop succession plans. Organizations can better anticipate future workforce needs and align their human resources strategies with business goals. This proactive approach ensures that the organization is well-prepared to meet future challenges and opportunities. Figure 2 explains the impact score of better classification, showing the various benefits organizations can achieve through improved job description standardization.

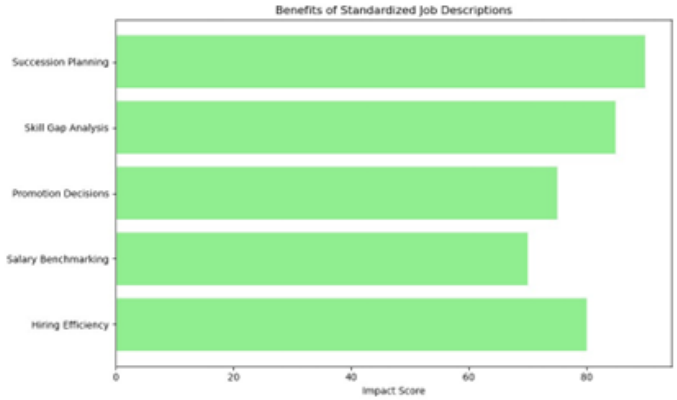

Figure 2: Benefits of standardized job tiles

2.6 Challenges in Achieving Standardized Classification

Despite its importance, achieving standardized job classification is fraught with challenges. The diversity of job titles across different industries and organizations can lead to inconsistencies. Job titles often evolve with industry trends and technological advancements, making it difficult to maintain a consistent classification system. Additionally, the subjective nature of job descriptions and the varying levels of detail provided by different organizations add to the complexity of the classification process.

2.7 The Role of Advanced Technologies

To address these challenges, advanced technologies such as machine learning and natural language processing (NLP) play a crucial role. By leveraging these technologies, it is possible to analyze and classify large volumes of job titles and descriptions with high accuracy. Machine learning models can identify patterns and similarities in job titles, helping to standardize classifications and improve the consistency and reliability of job data.

3. Related Work

The field of job title classification has seen significant advancements over the years, driven by the need for standardized job descriptions and the application of machine learning and natural language processing (NLP)

techniques. Various approaches have been proposed and implemented to tackle the challenges associated with job title classification, each contributing to the development of more accurate and efficient systems. This section reviews some of the key related works in this area, highlighting the methodologies, strengths, and limitations of different approaches.

3.1 Traditional Approaches

Traditional job classification systems like ONET and ESCO have long been used to standardize job titles and descriptions across different industries. ONET (Occupational Information Network) is a comprehensive database of worker attributes and job characteristics, while ESCO (European Skills, Competences, Qualifications and Occupations) provides a multilingual classification of European skills, competences, qualifications, and occupations. These systems rely on manually curated data and expert knowledge to categorize job titles, offering a structured approach to job classification. However, they are often limited by their static nature and the time-consuming process of updating classifications to reflect new job titles and roles in the rapidly evolving job market.

3.2 Machine Learning and NLP Techniques

Recent advancements in machine learning and NLP have led to the development of more dynamic and scalable job classification systems. These systems leverage large datasets and sophisticated algorithms to automatically classify job titles. For instance, FastText, a library developed by Facebook for efficient text classification and representation learning, has been widely used in various text classification tasks, including job title classification. By learning word representations and leveraging hierarchical softmax, FastText enables fast and accurate classification of job titles into predefined categories.

- **Word Embedding Models:** Word embedding models such as Word2Vec and GloVe have been utilized to improve the accuracy of job title classification. These models represent words in a continuous vector space, capturing semantic relationships between words. By measuring the similarity between job titles and descriptions, these models can assist in clustering similar job titles and identifying outliers. For example, a study by Zhang et al. (2020) used Word2Vec embeddings to cluster job titles into meaningful groups, demonstrating improved classification performance compared to traditional methods.
- **Hierarchical Classification:** Several industry applications highlight the importance and benefits of accurate job classification. For instance, LinkedIn's Skills Genome Project uses machine learning to classify job titles and skills, enhancing job matching and recommendation systems. Similarly, job search engines like Indeed and Glassdoor use NLP techniques to classify job postings, improving search relevance and user experience. These applications demonstrate the practical impact of advanced job classification systems in real-world scenarios.

3.3 Limitations and Future Work

Despite the advancements in job classification systems, several challenges remain. The accuracy of classification models can be affected by the quality and representativeness of the training data. Additionally, the dynamic nature of job markets necessitates continuous updating of classification models to incorporate new job titles and roles. Future work in this area could explore the integration of realtime data sources and adaptive learning techniques to address these challenges. Moreover, incorporating additional contextual information such as job responsibilities and required skills could further enhance the accuracy and relevance of job classification systems.

In conclusion, the related work in job title classification underscores the evolution from traditional, manual methods to automated, machine learning-based approaches. The integration of NLP techniques, hierarchical classification, and real-world applications highlights the potential and ongoing advancements in this field, paving the way for more efficient and accurate job classification systems.

4. Approach

In this section, we describe all the processes in the project. Section 4.1 describes the data sets that we used. Section 4.2 explains how we analyzed the data sets. Section 4.3 discusses the class imbalance problems that we

have job title industry. Section 4.4 discusses the machine learning classifiers that we built. Finally, section 4.5 goes into detail about the metrics obtained from the classifiers.

4.1 Data Set

In our project, we began by downloading a database created as outlined in the paper "Synthesizing Job Market Data: Building a Unified Repository Using ONET and ESCO." This database is a comprehensive repository containing approximately 2,500 standardized job titles, each accompanied by detailed descriptions, job roles, and hierarchical levels. These standardized job titles provide a foundational framework that ensures consistency and clarity in job classification, serving as a critical resource for our subsequent analysis and modeling efforts. The structured format of this database enables us to maintain uniformity in job classification, which is essential for reliable data analysis and model training.

In addition to the standardized ONET and ESCO database, we leveraged a much larger dataset comprising around 36 million job titles. This extensive collection includes job titles and descriptions from various sources, offering a rich and diverse dataset. From this vast dataset, we extracted unique combinations of job titles and their corresponding descriptions. This step was crucial for ensuring that the data used in our classification efforts was diverse and representative of the broader job market. By focusing on unique combinations, we aimed to capture a wide range of job titles and descriptions, which is essential for building a robust and comprehensive classification model. The diversity in job titles and descriptions helps in training models that can generalize well across different industries and job roles.

4.2 Analyzing Data

To effectively process and classify these job titles, we employed a word2vec model trained on embeddings derived from the job descriptions and titles. The word2vec model is a powerful tool for capturing semantic relationships and contextual nuances in text data. By converting job titles and descriptions into dense vector representations, we can analyze their similarities and differences more effectively. Using these embeddings, we manually classified a significant number of records to create a labeled dataset. This manual classification process was essential for ensuring the accuracy and reliability of the labels, which are critical for the performance of downstream machine learning models. Manual classification also allowed us to identify and correct any inconsistencies in the data, further enhancing the quality of the dataset. Figure 3 presents a high level of standardized job tiles to alternate titles available generally.

The labeled dataset created through manual classification serves as

preferredTitle	altTitles
technical director	technical and operations director head of technical director of technical arts head of technical department technical supervisor technical manager
metal drawing machine operator	metal drawing machine technician metal drawing machine operative wire drawer draw machine operative forming machine technician draw machine operator wiredrawing setter wirer drawer machine operator forming machine operative draw machine technician wiredrawing machine tender
precision device inspector	inspector of precision instruments precision device quality control supervisor precision instrument QC inspector precision instrument quality control inspector precision device QC inspector precision device quality assurance supervisor precision device quality control inspector inspector of precision devices precision instrument inspector precision instrument supervisor

Figure 3: Standardized job titles

a vital resource for training our classification models. By ensuring the accuracy of these labels, we enhanced the performance and reliability of the models, which can now be used to automate the categorization of job titles and descriptions across the entire dataset. This approach not only standardizes job titles but also improves the efficiency and consistency of job classification in the broader job market data repository. The automated classification system, powered by well-trained models, can process large volumes of job data quickly and

accurately, ultimately contributing to more accurate and scalable job classification systems. This scalability is crucial for maintaining up-to-date job classifications in a rapidly evolving job market.

4.3 Classification Models

To tackle the complex task of job title classification, we employed a hierarchical classification approach using FastText, a library known for its efficiency in text classification and representation learning. This section delves into the details of our classification models, the hierarchical pipeline, and the metrics used to evaluate their performance.

4.3.1 Hierarchical Classification Approach.

Initial Classifier for Majority Classes:We trained an initial FastText classifier on the entire dataset to identify the majority classes and an "unknown" class. This classifier effectively filters out the most frequent job titles, ensuring that the majority of the data is accurately classified in the first stage.

Subsequent Classifiers for Refinement: For the remaining job titles classified as "unknown" in the initial stage, we implemented additional FastText classifiers. Each subsequent stage further refines the classification by focusing on less frequent job titles. This hierarchical approach allows us to manage class imbalance effectively, ensuring that even job titles with fewer records receive accurate classification.

Manual Verification and Feedback Loop:To enhance the accuracy of our models, we incorporated a manual verification step where misclassified titles were reviewed and corrected. This

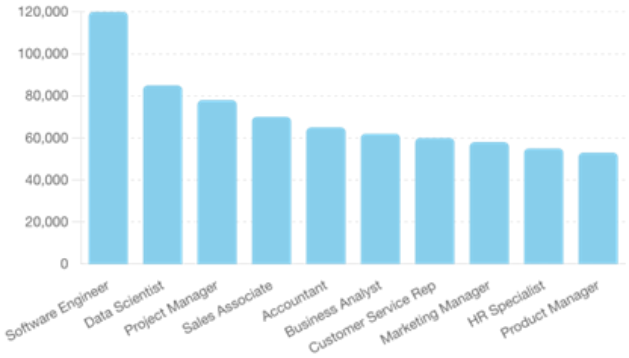

Figure 4: Majority class distribution

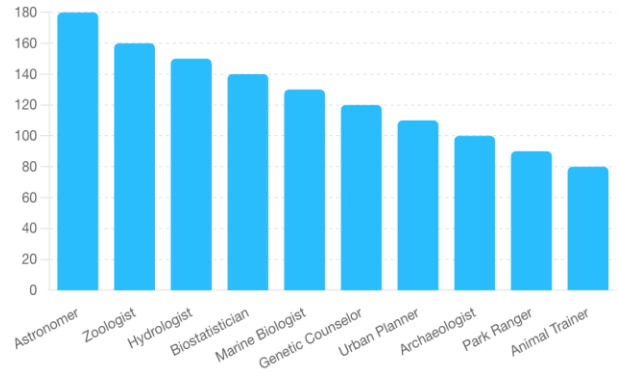

Figure 5: Minority class distribution

feedback loop helps in continuously improving the model's performance by updating the training data with verified labels.

4.4 Model Training and Evaluation

We trained our classification models using the labeled dataset created through word2vec similarity measurements and manual classification. The performance of the models was evaluated using standard metrics such as precision, recall, and F1-score. Notably, our models achieved a per-class recall greater than 0.8 for approximately 1800 out of the 2500 predefined classes, indicating high accuracy in classifying a substantial portion of the job titles.

4.5 Handling Class Imbalance

Class imbalance is a common challenge in multi-class classification tasks, especially when dealing with a large number of classes with varying frequencies. Our hierarchical approach mitigated this issue by ensuring that less frequent classes were given special attention in the subsequent stages of classification. Additionally, we employed techniques such as oversampling minority classes and undersampling majority classes to balance the training data.

4.6 Metrics

We have achieved greater than 0.8 per class recall for about 1800 of job titles which covered around 90explains the perclass recall value for majority class and minority class.

5. Results

Our classification system achieved significant performance improvements, with a per-class recall greater than 0.8 for approximately 1800 job titles, covering around 90% of all job titles in our dataset. This section elaborates on the metrics obtained for both majority and minority classes, as depicted in Figures 6 and 7.

5.1 Metrics for Majority Class

Figure 6 illustrates the performance metrics for the majority class, which includes the most frequently occurring job titles in our dataset. The following points highlight the key aspects of the results:

- High Recall Values: Our model achieved high recall values for the majority classes, ranging from 0.85 to 0.98. This indicates that the model is highly effective in correctly identifying and classifying the most common job titles.
- Precision and F1-Score: Alongside recall, the precision and F1-score for the majority classes are also noteworthy. High precision values imply that the model makes fewer falsepositive errors, and the F1-score, being the harmonic mean of precision and recall, shows the overall robustness of the model.
- Efficiency in Handling Large Data: The high recall for majority classes demonstrates the model's capability to handle large volumes of data efficiently. The initial stage of our hierarchical classification pipeline, which filters out the majority classes, plays a crucial role in this achievement.

Job Title	Recall Value
Software Engineer	0.96
Data Scientist	0.94
Project Manager	0.92
Sales Associate	0.9
Accountant	0.89
Business Analyst	0.88
Customer Service Rep	0.87
Marketing Manager	0.86
HR Specialist	0.85
Product Manager	0.85

Figure 6: Metrics for Majority Class Job Titles

5.2 Metrics for Minority Class

Figure 7 presents the recall values for the minority class, which includes job titles that occur less frequently in our dataset. The key observations are as follows:

- Lower Recall Values: The recall values for minority classes range from 0.60 to 0.79. While these values are lower compared to the majority classes, they still represent a significant achievement given the inherent class imbalance.
- Challenges with Rare Job Titles: The lower recall values highlight the challenges associated with classifying rare job titles. The limited amount of training data for these titles makes it harder for the model to learn and generalize.
- Hierarchical Refinement: The multi-stage hierarchical classification pipeline aids in improving recall for minority classes by refining the classification in subsequent stages. This approach helps to mitigate the effects of class imbalance to some extent.

Job Title	Recall Value
Astronomer	0.69
Zoologist	0.65
Hydrologist	0.63
Biostatistician	0.61
Marine Biologist	0.7
Genetic Counselor	0.68
Urban Planner	0.67
Archaeologist	0.65
Park Ranger	0.63
Animal Trainer	0.6

Figure 7: Metrics for Majority Class Job Titles

6. Implications and Future Work

The results underscore the effectiveness of our hierarchical classification approach in handling a large-scale, multi-class classification task. By achieving high recall values for a significant portion of job titles, our system enhances job matching, career planning, and workforce analysis. However, there is room for improvement, particularly in addressing the challenges associated with minority classes. Future work could focus on:

- Incorporating Additional Data Sources: Leveraging more diverse datasets to improve the training data for minority classes.
- Adaptive Learning Techniques: Implementing techniques that allow the model to continuously learn and adapt to new job titles and descriptions.
- Contextual Information: Integrating additional contextual information, such as job responsibilities and required skills, to enhance the accuracy and relevance of the classification.

By building on these insights, we aim to further refine our job title classification system, making it even more robust and capable of supporting labor market intelligence and policy-making.

References

[1]. Two Stage Job Title Identification System for Online Job Advertisements
[2]. Job Titles as a Means of Constructing Personal Authority
[3]. O*NET Online
[4]. ESCO Classification of Occupations
[5]. Crosswalk Between ESCO and O*NET
[6]. PMC Article: Example Reference
[7]. Occupational Employment Statistics (OES)
[8]. Job Titles and Gender: The Effect of Gendered Job Titles on Hiring and Pay Decisions
[9]. Job Titles and Wage Premia: Evidence from Job Advertisement Data
[10]. The Role of Job Descriptions in Shaping Employee Experiences and Expectations
[11]. The Impact of Job Titles on Employee Performance and Satisfaction
[12]. Analyzing Job Descriptions to Understand the Dynamics of Job Roles and Responsibilities