BULGARIAN ACADEMY OF SCIENCES

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Emiliano Maksim Mankolli

OPTIMIZATION METHODS FOR MACHINE LEARNING APPLICATIONS

A B S T R A C T

of thesis for awarding educational and scientific degree PhD

Doctoral Program: „Informatics“
Professional area: „4.6. Informatics and Computer Science“

Scientific supervisor: Professor Vassil Guliashki, PhD

Sofia, 2023

The thesis contains 130 pages, 17 figures, 5 tables and 131 bibliography sources.

Introduction

In the dynamic landscape of the recruitment industry, the integration of Machine Learning (ML) and Natural Language Processing (NLP) is pivotal. This thesis delves into the applications of ML and NLP, focusing on enhancing and streamlining the hiring process. An exhaustive review of the existing literature has been conducted to pinpoint the current strengths, limitations, and emerging gaps in the field. It is proposed to utilize ML and NLP techniques to elevate the quality of candidate selection, aiming for efficiency and reduced intervention, thereby reducing the time and costs associated with recruitment.

Central to this thesis are novel algorithms and models that not only refine the recruitment process, particularly in candidate screening and job candidate matching, but also emphasize computational efficiency and reduced algorithmic complexity. This aspect is vital, as it addresses one of the major challenges in applying ML and NLP in large-scale recruitment processes: the often prohibitive computational costs and extended execution periods.

The initial filtration process in candidate screening was enhanced using ML and NLP techniques. Language models are deployed to shift through vast numbers of candidate profiles and resumes to extract salient information. These were then processed by downstream ML algorithms to identify the most suitable candidates. Furthermore, job matching was scrutinized by utilizing ML and NLP to align job requirements with candidate capabilities more accurately. Semantic analysis, information retrieval, and recommendation systems constitute the backbone of the proposed framework.

This thesis reveals the transformative potential of ML and NLP in the recruitment sector. Recent advancements in machine learning can expedite crucial hiring stages, including candidate screening and job matching, enabling organizations to optimize operations, diminish biases, and enhance recruitment outcomes. The findings contribute significant methodological advancements and offer practical insights for HR professionals and industry practitioners, laying a robust foundation for future research on recruitment technology optimization.

This study is set against the backdrop of AI's profound impact across various disciplines, with ML emerging as a powerful tool in creating intelligent systems that mimic human intelligence. The research initially focused on a comprehensive review of the existing literature in the field. The optimization of recruitment processes through ML has shown remarkable

efficacy. Recruitment, a vital component of human resource management, has traditionally been labor-intensive and prone to bias. The advent of ML and NLP heralds a new era of automation and optimization in various recruitment phases.

The realization of AI and ML in the future of recruitment motivated this study. The application of ML algorithms in recruitment is proposed to revolutionize the industry and enhance efficiency and impartiality. The hiring process has been inefficient in the past because resumes had to be screened by hand and decisions were made based on bias. ML and NLP are changing this by promising more streamlined processes, fewer biases, and better candidate selection.

NLP, which is integral to ML, plays a pivotal role in this transformation. NLP algorithms are capable of parsing job descriptions, applications, and candidate profiles, thereby allowing for automated resume screening and skill-based job matching. In addition, NLP facilitates the creation of intelligent chatbots and virtual assistants, enhancing candidate engagement and satisfaction.

Moreover, with ML's expanding applications across domains, a key challenge lies in optimizing processes and algorithms to be more time- and resource-efficient. This research contributes to this endeavor in the recruitment industry by leveraging data analysis and automation to accelerate the recruitment cycle and enable more effective resource allocation by HR professionals.

In conclusion, this thesis acknowledges the sweeping changes in the labor market influenced by globalization, the COVID-19 pandemic, and the advent of new professions. The increased volume of job applicants necessitates methodologies that can handle large data volumes, a domain in which ML and NLP excel. Numerous companies are now adopting AI to streamline their recruitment processes, marking a significant shift in the industry.

Around 24% of businesses have adopted artificial intelligence as a method of talent acquisition, according to a survey by The Sage Group in 2020. According to a recent study, the majority of managers *(56%)* expressed their intention to adopt automated technologies in the upcoming year [1]. According to a reliable source, the projected value of the AI platform market is expected to reach $29 billion by the year 2019, with a subsequent increase of $52 billion by 2024 [2].

While there is a significant prevalence of AI utilization in the field of recruiting, candidates often have a considerable degree of skepticism regarding the efficacy of their evaluation using

intelligent systems. In a survey published in 2019, the focus was on the influence of AI in the recruitment sector. The following responses that the candidates gave when asked whether they preferred to have a human interviewer or a computer interview introduced a novel method of hiring.

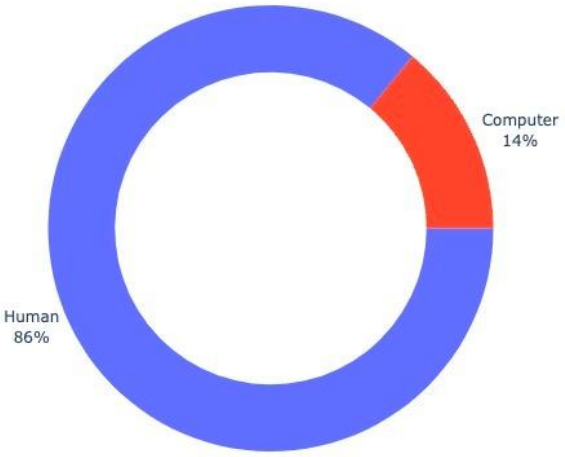

Figure 1. Preference percentage for the interviewer

This suggests that the level of skepticism among the applicants regarding the effectiveness of these algorithms is substantial. Subsequently, the candidates were asked, "Have you ever submitted an application for employment and not received any response?" The graph below shows the responses provided by the candidates.

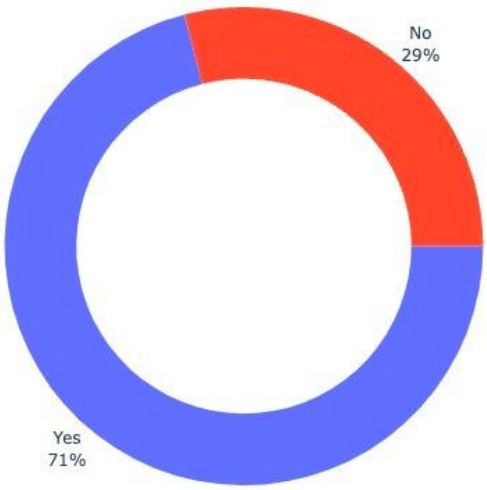

Figure 2. Percentage of companies that respond to applicants

The observed outcome can be attributed to the challenges faced by recruiters or human resources (HR) firms in effectively managing the data of all potential candidates. Based on the

aforementioned findings, it can be inferred that the utilization of artificial intelligence (AI) in the field of recruitment has transitioned from optional to imperative.

The candidate selection process involves evaluating the alignment between a candidate's CV elements, such as job title, years of experience, skills, and job criteria. This thesis examined an algorithm that utilizes machine learning techniques to identify similarities between two job titles. The primary objective is to ascertain job titles that can be deemed viable replacements for vacant positions.

In pursuit of these objectives, this dissertation consists of four interconnected studies that employ ML and NLP techniques to optimize particular phases of the recruitment process [3] [4] [5] [6].

These studies aimed to demonstrate the transformative potential of machine learning and natural language processing techniques for recruitment. By utilizing these technologies, we aim to provide HR professionals with invaluable insights and advance the recruitment industry. The idea that ML and AI can improve people's lives and encourage fair competition in the recruitment industry is what drives our research. A future is envisioned where recruitment processes are streamlined, biases are minimized, and candidates are accurately matched with opportunities that align with their talents and aspirations.

This thesis endeavors to illuminate the convergence of machine learning, natural language processing, and recruitment. Its goal is to optimize recruitment processes, diminish biases, and enhance overall outcomes through the application of these technologies. The research is directed towards providing HR professionals with essential tools and knowledge for making informed decisions, thereby fostering a more equitable and effective recruitment ecosystem. Ultimately, by harnessing the capabilities of ML and NLP, this thesis aims to refine recruitment practices and facilitate meaningful connections between candidates and organizations.

Chapter 1

Survey of the State of the Art: Machine Learning, Natural Language Processing, and Optimization

This chapter provides a comprehensive survey of the latest advancements in machine learning (ML) and natural language processing (NLP). These fields are at the forefront of technological evolution, especially within the domain of recruitment. The aim is to provide a detailed overview of the current state-of-the-art techniques, highlighting the most innovative and impactful developments in these areas.

The focus is on an in-depth examination of contemporary ML and NLP methods, which are considered benchmarks of excellence.

1.1 Natural Language Processing (NLP) Methods and Applications

1.1.1 Introduction to Natural Language Processing

Natural Language Processing (NLP), or computational linguistics, is one of the most essential technologies of the twenty-first century. It is fundamentally interdisciplinary and founded on linguistics, computer science, and artificial intelligence. Machine learning is related to natural language processing in that it must "understand " natural language and perform complex tasks, such as language translation and question answering.

1.1.2 NLP Techniques for Text Preprocessing

The foundation lies in effective text preprocessing, a phase crucial for structuring and cleaning raw text data. Tokenization, Stemming, Lemmatization, Stopword Removal, Normalization, Noise Removal are some steps applied in preprocessing stage.

1.1.3 Text Representation and Feature Extraction in NLP

Machine comprehension of text requires effective representation and feature extraction. Some techniques are Bag-of-words (BoW), TF-IDF, Word Embeddings, n-grams, Part-of-speech (POS) tagging.

1.1.4 Named Entity Recognition and Entity Linking

Gleaning specific information from vast textual data involves the recognition and linking of entities: Named Entity Recognition (NER), Entity Disambiguation, Entity Linking.

1.1.5 Sentiment Analysis and Opinion Mining

In the age of online reviews and social media, extracting sentiments and opinions from texts is invaluable. Polarity Detection, Emotion Detection, Aspect-based Sentiment Analysis, Opinion Extraction are some aspects of this.

1.1.6 Text Classification and Document Clustering in NLP

Organizing and categorizing vast volumes of text is instrumental for effective data analysis.

1.1.7 NLP Applications in Recruitment

Modern recruitment strategies harness the power of NLP for efficiency and precision.

1.2 Machine Learning (ML) Techniques

1.2.1 Introduction to Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence concerned with the development of algorithms and models that enable unprogrammed computers to learn patterns and make predictions or decisions based on data. ML algorithms have been widely implemented in numerous domains, including Natural Language Processing (NLP), where they have substantially contributed to the field's advancement.

1.2.2 Supervised Learning Algorithms

Supervised learning algorithms construct a generalizable model from input-output pairings. Various tasks, including text classification, sentiment analysis, named entity recognition, and machine translation, have utilized supervised learning algorithms in the context of NLP.

1.2.3 Unsupervised Learning Algorithms

When data is unlabeled, unsupervised learning algorithms are used to uncover latent patterns or structures within the data.

1.2.4 Reinforcement Learning Algorithms

Reinforcement learning (RL) is a learning paradigm in which an agent interacts with its environment and learns to maximize a cumulative reward signal by taking action [19].

1.2.5 Semi-Supervised and Active Learning Techniques

Semi-supervised learning techniques enhance the performance of NLP models by utilizing both labeled and unlabeled data.

1.2.6 Deep Learning and Neural Networks

Deep Learning, powered by neural networks, has revolutionized NLP by attaining cutting-edge performance on a variety of tasks. Neural networks consist of interconnected neurons that mimic the structure and functions of the human brain.

1.2.7 Transfer Learning and Ensemble Methods

Transmission learning permits the transmission of knowledge from one task or domain to another related task or domain.

1.3 Important Machine Learning Models

1.3.1 Supervised ML models

A. Sequential models

- Deep Learning
- Conditional random field (CRF) model
- Hidden Markov Model (HMM)
- Maximum entropy model (MaxEnt)
- k-Nearest Neighbors (k-NN)

B. Non-sequential models

- Naive Bayes
- Decision trees (DT)
- Support vector machines (SVM)

1.3.2 Unsupervised ML models

- Clustering
- Vector quantization

1.3.3 Reinforcement learning

- Proximal Policy Optimization (PPO)

1.3.4 Semi-supervised ML models

- Bootstrapping

1.3.5 Deep Learning and Neural Networks

- Transformer Model (BERT)

1.3.6 Transfer Learning and Ensemble Methods

- Stacked Ensemble Model

1.4 Optimization Problems in Machine Learning

In the machine learning process, an optimization problem is formulated, and the extremum (maximum or minimum) of an objective function is sought. The objective function is the objective or criterion that the model seeks to optimize. During the first step of any machine learning method, a good model for the job and a reasonable objective function that matches the goal are chosen.

1.4.1 Supervised learning optimization problems

The goal of supervised learning is to create an ideal mapping function, $f(x)$, that accurately predicts the labels given input features. The optimization problem consists of minimizing the loss function of the training samples, which quantifies the difference between the predicted and actual labels.

1.4.2 Semi-supervised learning optimization problems

The SSL bridges the gap between supervised and unsupervised learning. It utilizes both labeled and unlabeled data during the training process, making it useful in situations in which

acquiring labeled data is costly or time-consuming. SSL can be used in a variety of applications, including classification, clustering, regression, and dimensionality reduction.

1.4.3 Unsupervised learning optimization problems

Clustering algorithms [103] aim to group samples into multiple clusters based on their similarities. The objective is to minimize the differences between samples within the same cluster while maximizing the dissimilarities between samples in different clusters.

1.4.4 Trends of ML and NLP

Owing to these developments, NLP applications have become more accessible, efficient, and cost-effective. Language models that have already been trained, such as ULMFiT, CoVe, ELMo, OpenAI GPT, BERT, OpenAI GPT-4, XLNet, RoBERTa, and ALBERT, have revolutionized the NLP landscape.

❖ The Goal of the Thesis is:

- ★ To develop and refine innovative machine learning (ML) and natural language processing (NLP) algorithms to revolutionize the recruitment industry by focusing on enhancing the efficiency and effectiveness of the recruitment process. This goal includes the development of new algorithms for optimized candidate screening, advanced job-candidate matching, and streamlined initial filtration processes, while maintaining computational efficiency and algorithmic simplicity. The aim is to provide a cutting-edge tool for recruitment companies to transform industry practices by leveraging semantic analysis, information retrieval, and recommendation systems.

The primary research objectives addressed in this thesis can be divided into two categories:

- The first main goal is to give an overview of the most up-to-date machine learning (ML) and natural language processing (NLP) techniques, as well as different ways to make them work better.
- The second and most important objective is the presentation of algorithms that integrate ML and NLP techniques to optimize the recruitment time and resources.

This study contributes to some specific nodes of the recruitment process that are often considered bottle necks.

The tasks for which this research provides solutions can be formulated as follows.

Task 1: to present state-of-the-art of ML and NLP methods with focus in application of them in recruitment industry.

Task 2: to present optimization algorithm for reducing the pool of candidates based on job title similarity and industry relevance by using hybrid ML techniques.

Task 3: to present advanced optimization algorithm for reducing the pool of candidates based on job title similarity and industry relevance by using cutting edge ML methods.

Task 4: to identify most important features for predicting job success for candidates.

Task 5: to create a prediction model for job success that integrates both quantitative and textual data.

Task 6: to address problems and challenges related to the application of AI in recruitment and offering an advanced solution for facing these challenges.

Chapter 2

Streamlining the Candidate Pool: Optimization of Accuracy and Efficiency for Job Title Similarity

In this chapter, the focus is on revolutionizing the candidate selection process in recruitment through the innovative application of ML and NLP. This chapter introduces a novel hybrid method that synergistically combines the k-nearest neighbor (kNN) algorithm with support vector machines (SVM) to refine the candidate pool. This approach is centered on leveraging job title similarity and industry relevance to categorize candidates efficiently and accurately. By exploring this hybrid method, this chapter aims to demonstrate how combining these distinct ML techniques can optimize the recruitment process, enhance the precision of candidate job matching, and significantly streamline the pool of candidates based on relevant job titles and industry contexts. This approach promises to be a significant step in making recruitment processes more efficient and effective, offering a detailed exploration of the implementation of the method and its potential impact on the recruitment industry.

- Problems with Varying Job Titles
- Word2Vec, BERT, and KNN: Embracing NLP's Power
- Increasing Recruitment Effectiveness
- Summary

Using advanced NLP techniques, industry embeddings, and classification models, this chapter describes a ground-breaking plan to change the way candidates are chosen and how quickly they are hired. Using Word2Vec, BERT, and KNN, it is determined the semantic relationships between job titles and optimized candidate selections based on industry similarity. Our research revolutionizes the recruitment process, ushering in an era of data-decision-making, and ensuring a perfect match between job titles and candidate abilities. With this comprehensive strategy, it intends to streamline the candidate pool, improve the quality of candidate evaluation, and create a streamlined and effective recruitment process that connects the most qualified candidates with ideal job openings.

2.1 Dataset Description and Relevance

Our datasets include the names of employment titles and industries, as well as their corresponding brief descriptions obtained through web scraping. These datasets serve as the basis for our research, allowing us to extract valuable insights using the sophisticated NLP techniques Word2Vec and BERT for effective candidate selection.

Job Title Dataset:

This dataset comprised approximately 80,000 unique job titles.

Industry Dataset:

In addition to the dataset of job titles, here is compiled information on approximately 700 industries.

Relevance of datasets:

The datasets are of utmost significance to our research because they provide the foundation for identifying job title similarities and optimizing the candidate pool. The primary characteristics that highlight the significance of our datasets are as follows:

1. Semantic context and skill correlations
2. Industry-Specific Insights
3. Word2Vec Embeddings
4. BERT Embeddings
5. Data-driven Decision-Making

2.2 Candidate Pool Streamlining Algorithm: Harnessing Job Title Similarity for Efficient Recruitment

One-third of talent acquisition professionals spend more than 20 hours sourcing candidates for a single position [108]. Talent acquisition professionals spend more than 30 percent of their workweeks sourcing candidates for a single position. In addition, the average American business spends $4,000 on engaging a new employee [109] and [110]. These statistics emphasize the need for more effective and economical recruitment procedures.

The traditional recruitment process has a negative impact on a number of economic factors for both candidates and employing companies. There are bottlenecks when working with

hiring managers, with 50 percent of recruiters encountering problems in moving candidates through the recruiting process and forty-four percent citing hiring managers' resume reviews as reasons for slowing the process [111]. 67% of the recruiters [112] cite the scarcity of competent and highly qualified candidates as their greatest hiring obstacle.

2.2.1 k-NN and SVM methods

A. k-Nearest Neighbors (k-NN)

B. SVM: Support Vector Machines

2.2.2 Optimization Method I: Maximizing the precision of job similarity

This section describes a proposal for the hybrid application of the SVM and k-NN methods to locate similar job titles.

The primary characteristic of a job title is its definition, which broadly describes what it entails. Accuracy issues begin to appear when the definition is not explicitly stated.

A specific job title is also associated with a variety of duties and responsibilities that the candidate must perform in the course of his daily activities. Some job titles clearly differentiate duties and responsibilities, making it simple to determine whether they are interchangeable. This similarity renders the "area" of responsibilities and duties a "gray area." This implies that a substantial portion of the duties and responsibilities of the two distinct job titles can be the same. In other words, they are interchangeable job titles.

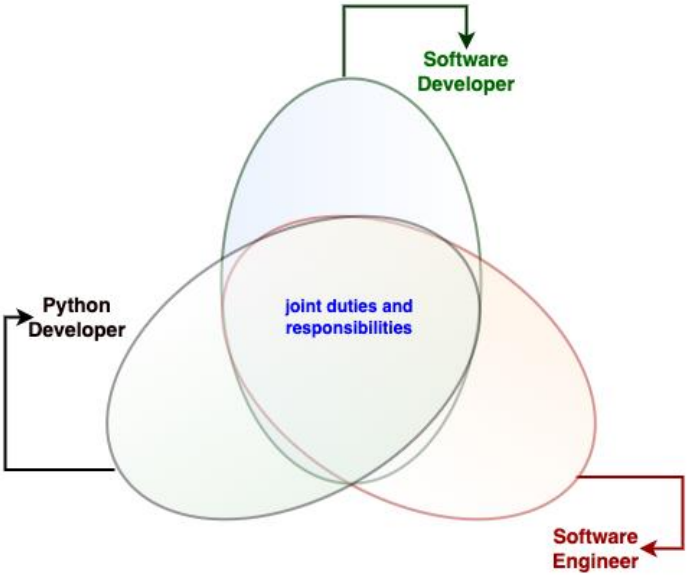

Figure 3. Example of similar job titles

The hybrid method seeks to improve the process of locating job titles similar to a given title. This is accomplished by integrating two crucial factors: optimizing execution time and memory allocation and enhancing the precision of similarity measurements. The hybrid application combines two machine learning techniques: Support Vector Machines (SVM) and k-nearest neighbors (kNN).

Optimization of Execution Time and Memory Allocation: During the computation of job title similarity, the hybrid method optimizes the execution time and memory allocation.

Improving Similarity Accuracy: Accurate measurement of job title similarity is essential for the success of the hybrid method.

The hybrid method incorporates the advantages of SVM and k-NN to provide a more exhaustive and accurate representation of job-title similarity. The hybrid approach is better at matching job titles because it takes the best parts of both methods and puts them together. This is especially true when SVM or k-NN alone might not work.

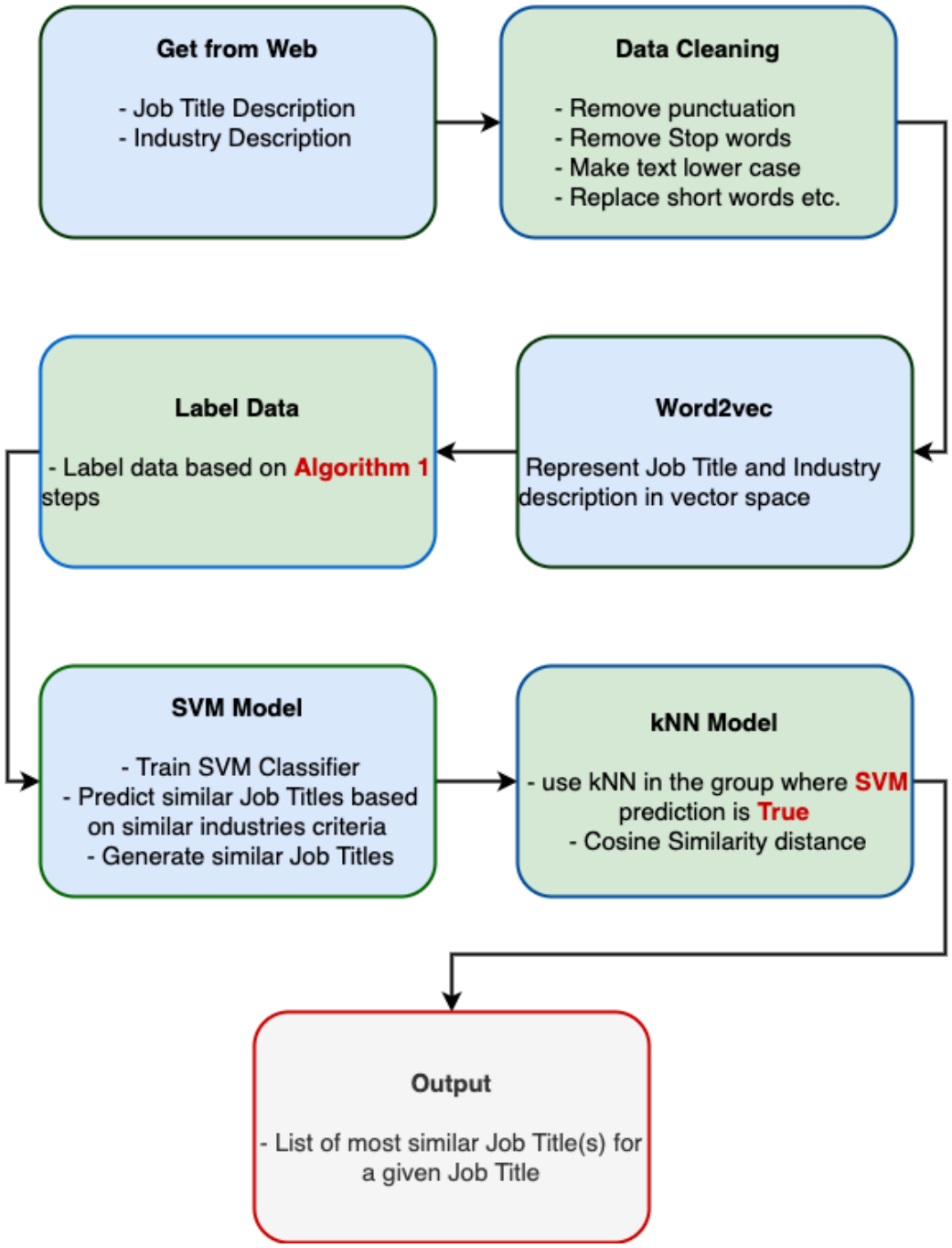

Figure 4. Optimization Method I: Based on Hybrid Approach I

The data utilized in the application's initial phase consisted of job titles and industry descriptions.

These data were preprocessed to generate a final dataset, which was represented by an embedding vector containing 100 entries for each occupation and industry. As its name suggests, Word2vec represents each distinct word with a vector, which is a specific list of integers. [124]

Two methods are available for obtaining embedding: Skip Gram and Common Bag Of Words (CBOW). Skip-gram Word2Vec is a computing architecture for word embedding.

Instead of using surrounding words to predict the center word, as CBow Word2Vec does, skip-gram uses only the center word. Word2Vec predicts the surrounding syllables based on a central word.

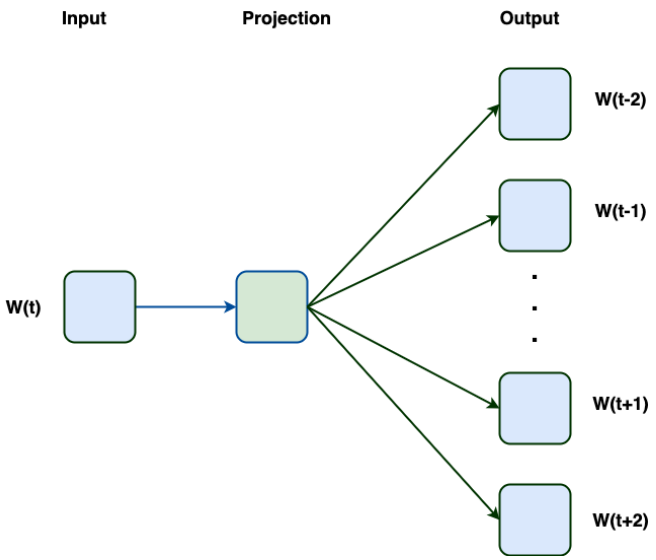

Figure 5. Skip-gram flow

The Skip-gram objective function produces the following objective function [124] by adding the log probabilities of n words to the left and right of the target word wt.

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-n\leq i\leq n \\ i\neq 0}} \log P(w_{j+1}|w_t),$$

where denotes the vector representation of each word.

Using vector similarity, it is defined the similarity between two vectors as [125]:

$$\text{Sim}(\overrightarrow{d},\overrightarrow{q}) = \frac{\overrightarrow{d}\,\overrightarrow{q}}{|\overrightarrow{d}||\overrightarrow{q}|} = \frac{\sum_{k=1}^{t} w_{q_k} w_{d_k}}{\sqrt{\sum_{k=1}^{t}(w_{q_k})^2}\sqrt{\sum_{k=1}^{t}(w_{d_k})^2}}$$

The most straightforward approach for identifying job titles similar to a given query involves calculating the cosine similarity between their corresponding vectors. However, this method presents two primary challenges that must be addressed for effective implementation.

- Quadratic Execution Time

One of the challenges is the execution time required for this approach. As the number of job titles increased, the execution time increased quadratically, resulting in longer processing

times. This is particularly problematic in scenarios where real-time or near-real-time results are necessary.

- Memory Usage Complexity

Memory usage complexity is another significant issue, in addition to execution time. Storing and managing a substantial number of vectors for all job titles requires a significant amount of memory. This can strain the computational resources of the system, potentially leading to memory bottlenecks and slower performance.

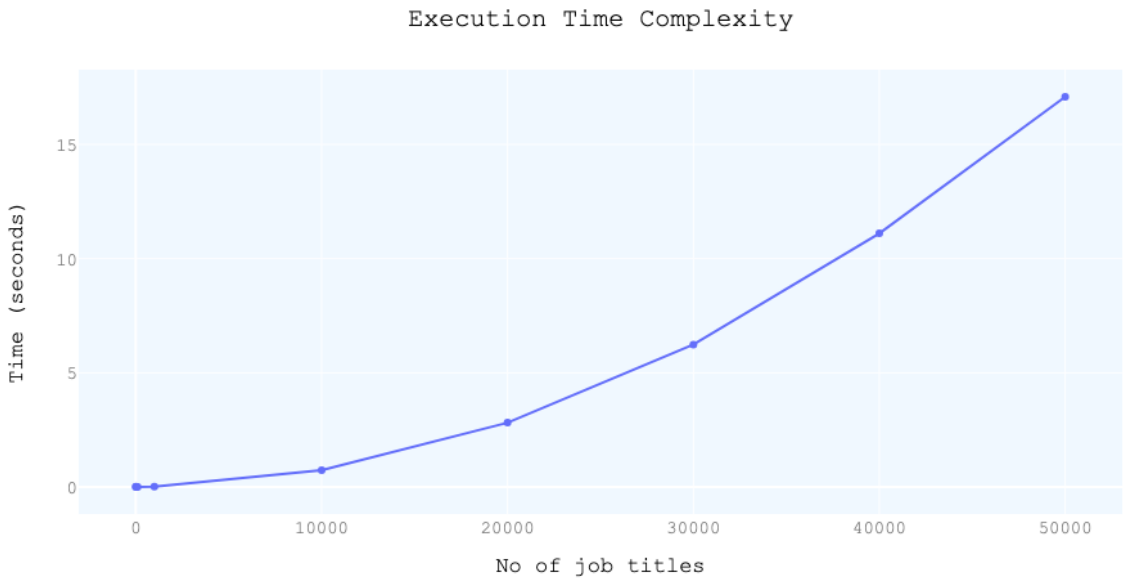

Figure 6. Execution time complexity

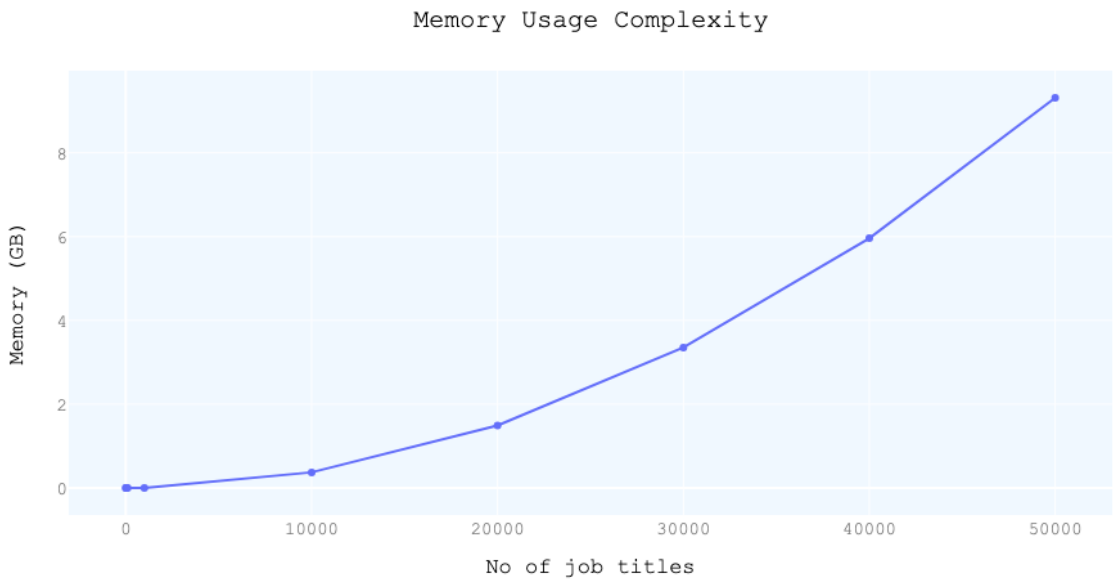

Figure 7. Memory Usage Complexity

*When measuring the cosine similarity (specifically, using the **Intel Core i9-10850K** processor), the computation time varied. It takes about **0.001 seconds** to calculate the similarity between a single pair of job titles, while this increases to **12.3 seconds** when processing 50,000 pairs of job titles.*

*With regard to memory utilization, the outcomes are equally noteworthy. Allocating approximately **120 B** of memory was required to determine the similarity between individual pairs of job titles. However, this memory allocation escalated significantly to approximately **19GB** when processing 50,000 pairs of job titles simultaneously.*

As depicted in the graphs, it becomes evident that as the number of job titles increases, the complexity of both execution time and memory usage also increases significantly. This observation underscores the suboptimal nature of this simplistic approach. Our solution, which entails the hybrid implementation of the SVM and k-NN methods, is designed to not only enhance speed but also to markedly alleviate memory usage intricacies.

With the vector embeddings for job titles and industries established via Word2vec, the subsequent steps involved employing the ensuing algorithm to construct a labeled dataset. This dataset serves as the foundation for training the SVM model.

Algorithm 1

Step 1: Calculate the similarity between Job Title and Industry using Cosine Similarity.
J = [Job title embedding]
I = [Industry embedding]
similarity = Cosine(J,I)
Step 2: Select n most similar industries for a job title.
Top_n (J)
Step 3: Find the joint industries of every pair of job titles by taking the intersection of the most similar industries for any pair of job titles.
joint_industry = $x \cap y$, where $x \in \text{Top\_n}(J_x)$, $y \in \text{Top\_n}(J_y)$
Step 4: Classify job titles industry similarity as 0 or 1 based on the threshold, where we compare ℓ, which indicates the number of joint industries between two job titles, with t, the criterion that determines how many joint industries must have two job titles to be called similar.

$$\text{industry\_similarity} = \begin{cases} 1, & \text{if } \ell \geq t \\ 0, & \text{otherwise} \end{cases}$$

where $\ell = \text{length}(x \cap y)$ and $t :=$ threshold.

An informed recommendation guided by heuristic models suggests that this criterion could encompass approximately *1-2%* of the total number of industries. This would result in the inclusion of approximately 5 to 7 shared industries from the upper echelons of the industry list corresponding to each job title.

Algorithm 1 is instrumental in creating a labeled dataset, wherein the embedding vectors for each job title constitute the features, whereas the classes generated in Step 4 serve as labels. Subsequently, Algorithm 2 was implemented to select the k most similar job titles for a given query job title.

Algorithm 2

Consider the job title J_1.
Step 1: Apply SVM on the embedding vector of J_1 and the embedding vectors of the remaining job titles of the dataset.
Step 2: Create a reduced dataset with the n job titles where SVM outputs 1 for the classification of industry_similarity.
Step 3: Apply k-NN method to the reduced dataset to find the k job titles most similar to J_1. The distance we use will be the cosine similarity.
Step 4: Show k job titles most similar to J_1.

2.2.3 Example of Results of Hybrid Approach I

To illustrate the efficacy of the innovative hybrid machine-learning approach, consider a scenario in which a recruiter aims to hire a data scientist while keeping the search broadened beyond this particular query. The application involves models trained on a dataset comprising *500 job titles* and *700 industries*.

As described in the first step of Algorithm 1, our procedure begins by calculating the degree of similarity between each job title and industry. The primary goal of this stage was to compile a list of relevant fields for each occupation. Step 2 of Algorithm 1 takes this list of related industries for each job title and uses it to determine the *10 industries* that are most similar. It is worth noting that the selection of the illustrative example's top industry count is the result of testing and heuristic methods. Diverse uses should account for no more than *2%* of the total number of industries.

As a result, there is a document in which every occupation is associated with a group of 10 relevant fields of work. The result of this effort was a symmetrical matrix with numeric values between 0 and 10.

$$\begin{bmatrix} 10 & 1 & 1 & \dots & 1 & 1 & 0 \\ 1 & 10 & 4 & \dots & 0 & 0 & 1 \\ 1 & 4 & 10 & \dots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 10 & 1 & 1 \\ 1 & 0 & 0 & \dots & 1 & 10 & 1 \\ 0 & 1 & 1 & \dots & 1 & 1 & 10 \end{bmatrix}$$

Moving on to Step 4 of Algorithm 1, initially is established the threshold as $t = 2$. It is crucial to emphasize that this threshold value, as exemplified in this demonstrative scenario, is an outcome of heuristic methodologies. Naturally, users retain the flexibility to employ distinct values based on the significance they assign to an industry in the context of candidate selection.

This threshold delineates a binary classification scheme, attributing values of either 0 or 1 to each pair of job titles. The outcome of this process results in a dataset that is suitably labeled. With this labeled dataset in hand, the SVM model is trained using the training data, subsequently subjecting it to testing using the designated testing segment of the data. The evaluation of the performance of the model is presented in the following sections.

Table 1. Model Classification Report (SVM)

THE EVALUATION OF THE MODEL

	precision	recall	f1-score	support
0	0.92	0.87	0.89	1243
1	0.87	0.92	0.89	1208
accuracy			0.89	2451
macro avg	0.89	0.89	0.89	2451
weighted avg	0.89	0.89	0.89	2451

The derived confusion matrix determined that the number of false positives and false negatives was relatively small.

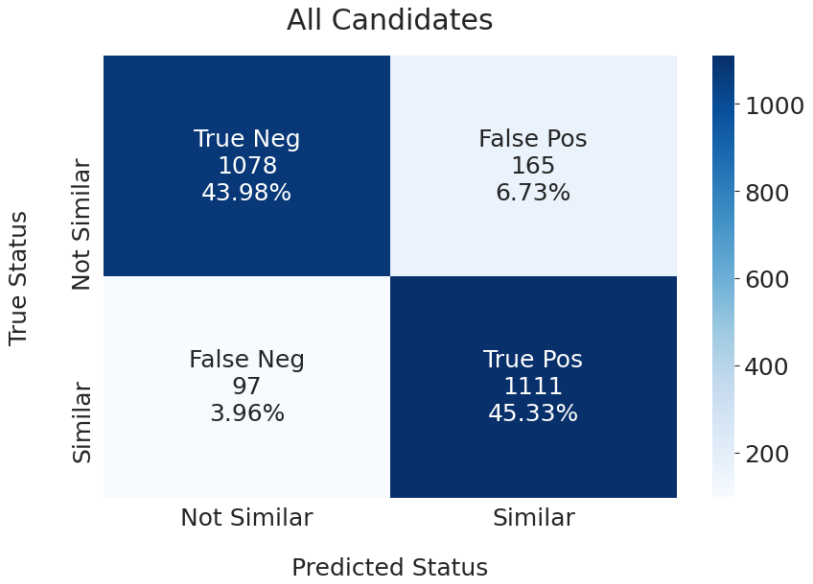

Figure 8. Confusion Matrix for Job Title similarity prediction - SVM model

The accuracy of the model was computed to be *89.3%*. As the culmination of this process, Algorithm 2 is then applied, yielding a listing of the 10 most similar job titles achieved through the hybrid methodology. For instance, considering the query job title "data scientist," Table 2 presents the results. If the objective is to recruit a candidate with the specified job title "data scientist," the algorithm would not only propose candidates explicitly holding this job title but also those with closely related job titles, as displayed in the second column of Table 2. The similarity score, which ranges from 0 to 1, quantifies how closely aligned the alternative job title is in relation to the query title.

Table 2. Results of substitute Job Titles for “Data Scientist”

DATA SCIENTIST MOST SIMILAR JOB TITLES

Query	Most Similar Job Titles	Similarity Score
data_scientist	data_scientist	1
	data_science_consultant	0.96577
	senior_data_scientist	0.95804
	lead_data_scientist	0.95755
	data_science_fellow	0.95331
	data_science_instructor	0.95038
	data_science_lead	0.94927
	data_science_mentor	0.93409
	machine_learning_engineer	0.93386
	statistical_modeler	0.92812

2.2.4 Conclusion and upcoming work

The outcomes illustrated in the table encapsulate the conclusive results obtained through the implementation of this hybrid methodology.

This novel hybrid application, which merges the SVM and k-NN methods, enhances the efficiency of the execution time and memory usage in locating similar job titles for a given query.

Looking ahead, further sections explore other high-performance methods for candidate selection. Beyond considering basic job title information, this process aims to incorporate additional features, such as specific job skills and the candidate's work history with previous companies.

2.3 Optimization Method II: For finding job title(s) best similar to a given job title.

2.3.1 Description of Machine Learning and Natural Language Processing Models

BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a groundbreaking NLP model that has redefined the landscape of language understanding tasks. Developed by Google AI researchers [25] , BERT leverages a transformer architecture, a deep neural network framework specifically designed to process sequential data, making it highly effective for tasks involving language understanding and generation.

BERT is pre-trained on an extensive corpus of text, including Wikipedia (~2.5 billion words) and the Book Corpus (~800 million words). This pre-training process is unsupervised and aimed at learning the statistical properties of language. During this process, BERT learns to represent words, phrases, or sentences in arrays of length 768.

XGBoost

XGBoost, standing for Extreme Gradient Boosting, is a distributed gradient boosting framework introduced with the primary intent of enhancing computational speed and model performance.

2.3.2 Description of Algorithms and Hybrid Approach II

The following diagram shows the flow of the proposed low-complexity hybrid method. In each step, the transformation from the inputs to the final output, which is the finding of a job title similar to a given job title, is described.

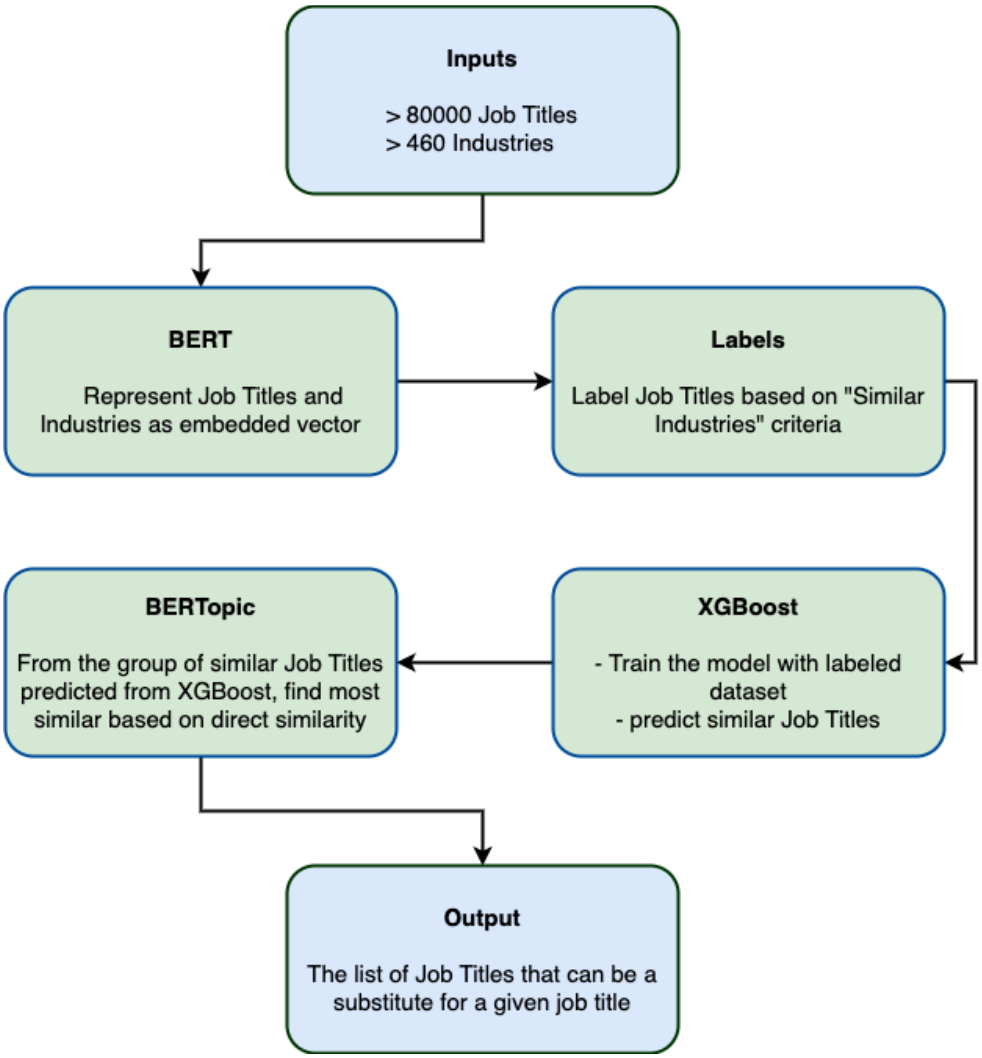

Figure 11. Optimization Method II: Based on Hybrid Approach II.

The data that this hybrid method receives as input is a finite list of 80,000 job titles and 480 industries.

Similar to the method explained in the previous section, the embedding of job title and industry was first performed using the BERT model. As a result, each job title and industry is represented as an array of length 768.

The "Labels" step shows how to define whether a pair of job titles are similar or not. This is performed using Algorithm 1 and Algorithm 2, as discussed in the previous section. The threshold used as the similarity criterion is 2, which means that two job titles are defined as similar if they have two common industries.
The industry is a decisive criterion for determining similarity. Among other things, a job title is related to knowledge of the domain where you work, its requirements, and its applications. To summarize, after this step, for each given job title, a list of similar job titles according to the criteria of common industries, is provided.

After the similarity label was defined, an XGBoost model was trained in the next step. This model can predict each new pair of job titles with a similarity class of 0 or 1. These job titles are similar based on common industries and are similar among them. After this step, for a given job title, our hybrid method generates a list of the top similar job titles that can serve as substitutes.
The figure below shows the Classification Report for the XGBoost model used for prediction. Compared to the SVM model used in the method presented in the previous section, the XGBoost model used in this hybrid method improved the classification accuracy by **2%.** There was also an improvement and balance in the prediction accuracy between classes.

Table 3. Model Classification Report (XGBoost)

	precision	recall	f1-score	support
0	0.91	0.90	0.91	14514
1	0.90	0.91	0.91	14638
accuracy			0.91	29152
macro avg	0.91	0.91	0.91	29152
weighted avg	0.91	0.91	0.91	29152

The confusion matrix of the XGBoost model confirmed the improvement in the prediction of the job title similarity class.

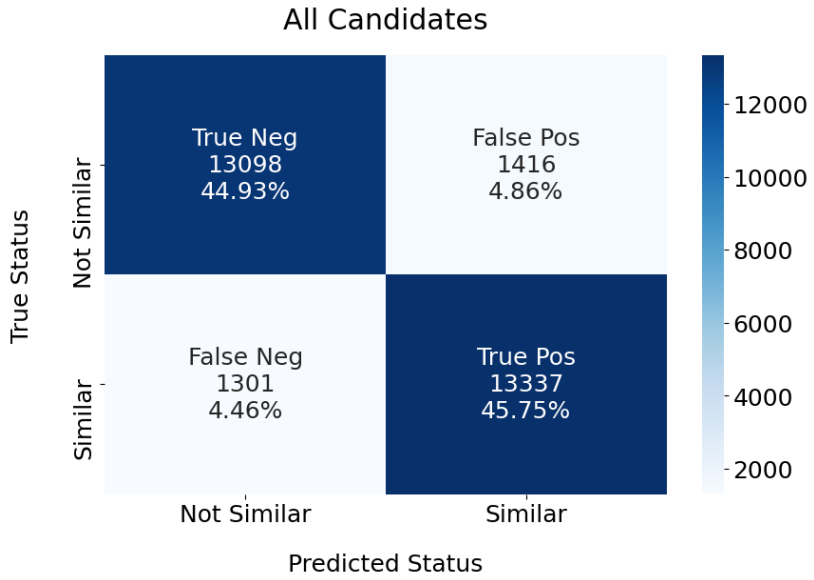

Figure 12. Confusion Matrix for Job Title similarity prediction - XGBoost model

The above results demonstrate that the XGBoost model predicts the job title similarity class with high accuracy. However, this is not enough, and for this reason, another step is included: BERTopic. This step aims to tighten the criteria for determining similarity. The following table shows the most similar job titles for "machine learning engineer ".

Table 4. Second hybrid approach results of substitute Job Titles for “Machine Learning Engineer”

Job Title	Similarity Score
machine_learning_engineer	1
software_engineer_machine_learning	0.95
senior_machine_learning_engineer	0.91
machine_learning_engineer_intern	0.91
machine_learning_scientist	0.91
machine_learning_researcher	0.90
machine_learning_research_assistant	0.89
artificial_intelligence_engineer	0.88
machine_learning_intern	0.86
machine_learning_research_intern	0.85

Chapter Summary and Conclusion

To summarize the findings of this section, the notable changes observed when using BERT as a transformer and XGBoost for predicting the similarity class are highlighted. It's important to recall that in the initial method, which shared a preprocessing step similar to the hybrid method discussed in this section, Word2vec was utilized as the transformer, and SVM was employed to predict the similarity class.

Some of the conclusions drawn from the hybrid approach are as follows.

- The use of machine learning techniques that are considered "state of the art" significantly improves the results of our hybrid method.
- *2%* improvement in the accuracy of job title similarity class prediction
- *11.6* times reduction in execution time and memory consumption, which makes this hybrid approach highly efficient for recruitment platforms that process big data.

For clarity, it is emphasised that the mentioned values are an average calculation. Job title similarity cannot be determined in the same way for all cases. For some job titles, a large number of other job titles can be found that can serve as substitutes, and for some other job titles, it is difficult to find substitutes.

Chapter 3

Predictive Models for Job Success. Optimization of candidate selection strategy

In the chapter, an advanced ML method is introduced for predicting job success, surpassing the efficiency of traditional manual approaches. This method is distinct in its utilization of BERT for embedding text and XGBoost for predicting job success. The integration of BERT enables a sophisticated understanding of textual data, thereby capturing the nuances of soft skills and other text-based candidate information. XGBoost complements this by offering a powerful and efficient predictive model known for its high performance in classification tasks. This chapter explores how the combination of these advanced techniques enhances a model's ability to accurately predict job success by incorporating a diverse array of data, including speech features. The synergistic use of BERT and XGBoost in this context illustrates a pioneering

approach in the recruitment field, showing how cutting-edge ML techniques can be leveraged to assess candidate potentials more comprehensively and accurately. Through this innovative methodology, this chapter contributes significantly to the field, demonstrating the potential of ML in transforming recruitment processes and improving the accuracy of job success predictions.

3.1 Introduction

Employment stands as one of the most crucial factors that directly impact the quality of our lives. Not only does it provide individuals with financial stability, but it also shapes their sense of purpose, identity, and overall well-being.

In the past decade, there has been a notable shift in the utilization of machine learning techniques for recruitment purposes.

One advantage of AI-powered recruitment tools compared to traditional recruitment methods is the reduced processing time.

3.1.1 Recruitment Evaluation Strategies

At the interview stage, recruiters carefully analyze candidate data to assess whether they possess the necessary skills, qualifications, and attributes to be a good fit for the job position. Recruiters typically consider various criteria to evaluate candidates effectively. Some of these criteria include:

1. Quality of hiring

2. Recruitment duration

3. Average employment costs:

$$\text{Average Employment Cost} = \frac{\text{recruitment cost}}{\text{number of hires}}$$

4. Retention Period/Job Turnover

3.1.2 Proposed solution

The utilization of machine learning techniques to predict the likelihood of candidate success, is proposed. In addition to experience and skills, is proposed including other textual and

speech data analyzed using natural language processing algorithms. Parameters such as the candidate's profile description, the job description, and other relevant text and speech data can significantly improve the accuracy of the predictive model.3.2 Theoretical Background and Empirical Findings

3.2.1 Human Capital Theory

A person's knowledge, experience, and skills are significant assets that increase their productivity and performance in the job market, according to the human capital theory that Gary Becker developed [127].

3.2.2 Social Cognitive Theory

Albert Bandura [128] developed the social cognitive theory, which emphasizes the interaction between people's cognition, behavior, and social environment. It asserts that individuals learn and develop their skills, self-efficacy, and motivation through observation, modeling, and social interaction.

3.2.3 Person-Environment Fit Theory

The person-environment fit theory underscores the relationship between an individual's attributes, the expectations of the job, and the broader organizational context.

3.2.4 Machine Learning Techniques in Predictive Modeling

Machine learning techniques have gained considerable attention in predictive modeling for job success. For example, the study conducted by M.L. Demircan and K. Aksaç [131] in the private banking sector in Turkey, analyzed a sample pool of 597 individuals. Their research utilized various machine learning models and achieved an accuracy rate of over 73% in predicting candidate success before employment commences.

3.3 The proposed Approach

The innovation in our methodology involves integrating additional data, including text and speech, to enhance the candidate's profile and achieve greater precision in predicting job success. BERT transforms the text data into a format that the machine learning model picked for the prediction stage can use. This is done during the preprocessing phase as well as the standard steps. After the preprocessing and feature engineering steps are finished, an XGBoost model that tells us whether the candidate will be successful or not, is trained.

3.3.1 Data Description

In the context of this study, is analyzed a dataset comprised of 648 individual entries. Each entry provides comprehensive details about an individual, encapsulating their job designation, associated industry, cumulative professional experience, predominant skills, descriptions of the most recent three workplaces or projects they've been involved with, and the average duration they've remained at a position. A particularly interesting component of our dataset are the nuanced insights sourced from their CV or video CV, which highlight elements of motivation, enthusiasm, and communication skills. Throughout this thesis, it si reffered to this compilation of information as the "candidate profile".

Additionally, the dataset features a "job success class," a binary metric signaling whether a candidate successfully navigated the project's probationary phase. While probationary durations can vary across projects, a consistent measure of success is a duration of 6 months. For the purposes of our modeling, the "job success class" is the specific target variable that the XGBoost Classifier model is trained to predict.

3.3.2 Pre-processing and Feature Engineering phase

In this section, the steps through which each input is prepared for use as a feature in the prediction model are described.

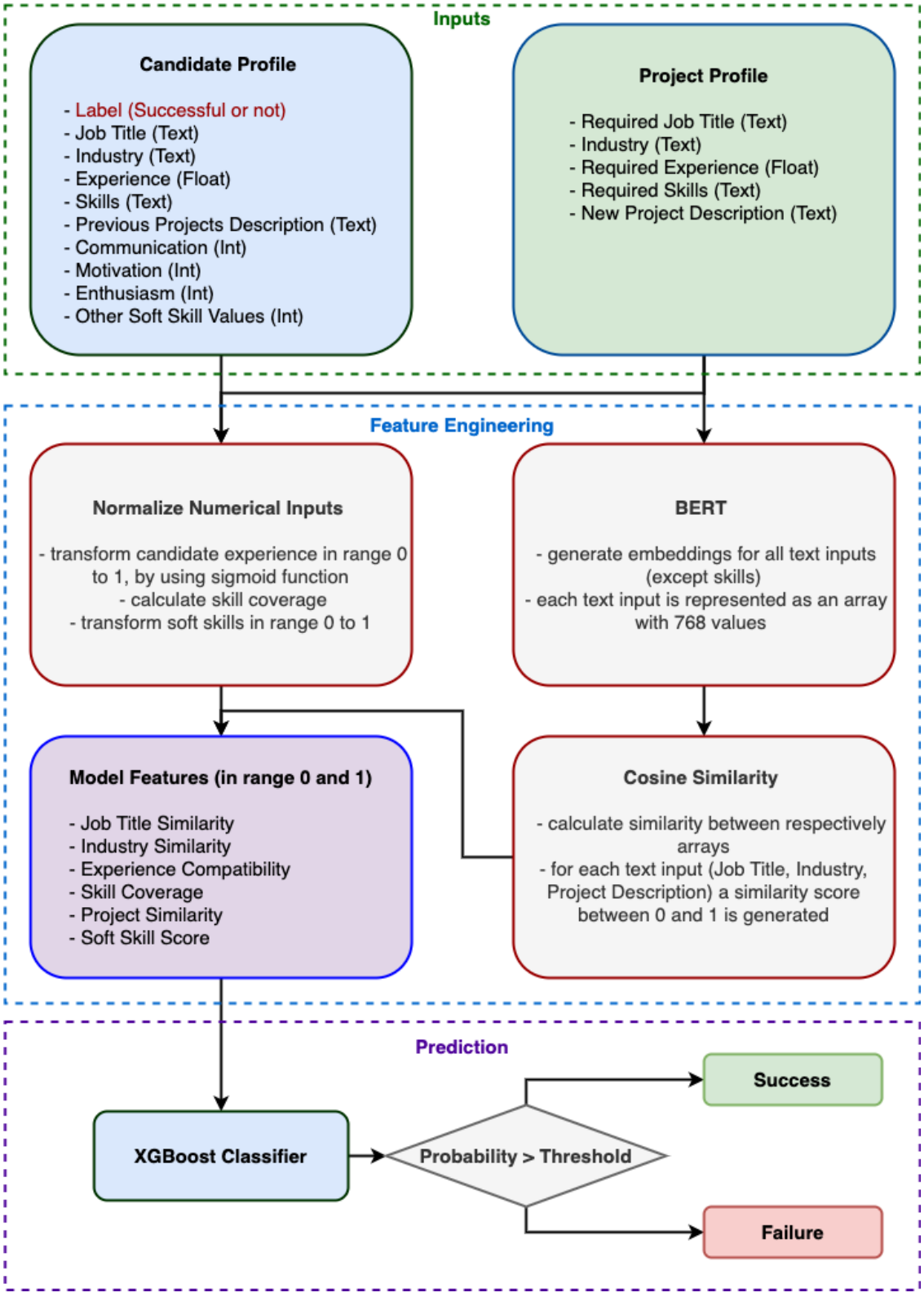

Figure 15. Job Success Prediction Model Flow

The job title, industry, and project description provided as textual data go through the same pre-processing procedure. The first step was embedding using the BERT model. The embedding process represented each data text, job title, industry, and project description in an array with a length of 768. Subsequently, the cosine distance was calculated for each pair of embedding arrays. This distance assigns each pair a similarity coefficient ranging from 0 to 1.

Usually, in a job vacancy, some skills that the candidate must have are given. In the feature engineering phase, skills are transformed into a numerical value from 0 to 1, which describes the percentage of skills required in the project profile possessed by the candidate. Therefore, this coefficient shows the intersection between the set of candidate skills and the set of skills required by the project, as shown in percentage terms.

The candidate's experience is one of the most important criteria that greatly affects the success or failure of a job position. In the feature engineering phase, a sigmoid function is created to transform the years of experience into a score between 0 and 1. This experience score takes small values (close to 0) if the candidate's experience (CE) is much less than required (RE) and high values (up to 1) if the candidate's experience is closer to what is required. The following formula shows how the Experience Score feature is calculated:

$$\text{Experience score} = \begin{cases} \dfrac{\text{CE}}{\text{CE} + e^{-c(\text{CE}-\text{RE})}}, & \text{if CE} \leq \text{RE} \\ 1, & \text{otherwise} \end{cases}$$

where c is a constant that represents aggressivity; in this case, $c = 0.5$.

The other inputs are collected through a structured procedure, and they are represented by integer numbers in the range from 1 to 10. In the feature engineering phase, these inputs are transformed by dividing them by their maximum, 10 in this case, resulting in features with values in the range 0 to 1.

3.4 Advantages of this Approach

The advantages of this approach are evident in at least two primary directions:

1. Efficiency

2. Accuracy

Machine learning approaches are known to be more efficient, specifically for time-consuming tasks such as the recruitment process. The proposed machine-learning approach makes a significant difference in reducing the time needed to process applicants, reducing costs, and many other benefits that arise in relation to classical recruitment methodologies.

In the treated experiment, the recruiters correctly predicted the success of the candidates in **84%** of the cases, while, after the interview stage, only **542** of the **648** selected candidates successfully passed the testing stage.

3.5 Model evaluation and result analysis

1. Inclusion of text and speech data

The proposed machine learning approach aimed to determine whether using text and speech data could improve the ability to select the correct candidates for a job.

❖ *Comparison of the Old Model and New Model*

Table 5 presents the main evaluation metrics used to compare the two models used in this experiment.

Old Model: This model considered only what is called "hard skills." These are measurable skills, such as a person's qualifications, years of experience, and technical knowledge.

New model: This newer approach not only considers hard skills but also integrates text and speech data. This would help assess "soft skills" such as communication skills, adaptability, and teamwork.

Table 5. Evaluation metrics of Old Model and New Model

Model	**Evaluation Metrics**				
		Precision		**Recall**	
	Accuracy	Class (1)	Class (0)	Class (1)	Class (0)
Old Model	0.88	0.97	0.59	0.88	0.87
New Model	0.93	0.98	0.72	0.93	0.91

❖ *Findings:* By including Text and Speech Data

- The overall model's accuracy improved by *5%*. This implies that the new model can predict suitable candidates *5%* better than the old model.
- The model's precision, which is a measure of how many selected candidates are actually suitable, increased by *1%*.
- The model's recall, indicating how many of the truly suitable candidates it could correctly identify, was improved by *5%*.

2. Machine learning vs. manual recruiting

Figure 17 shows the confusion matrix of the trained model with all features, which is referred to above as the "New Model".

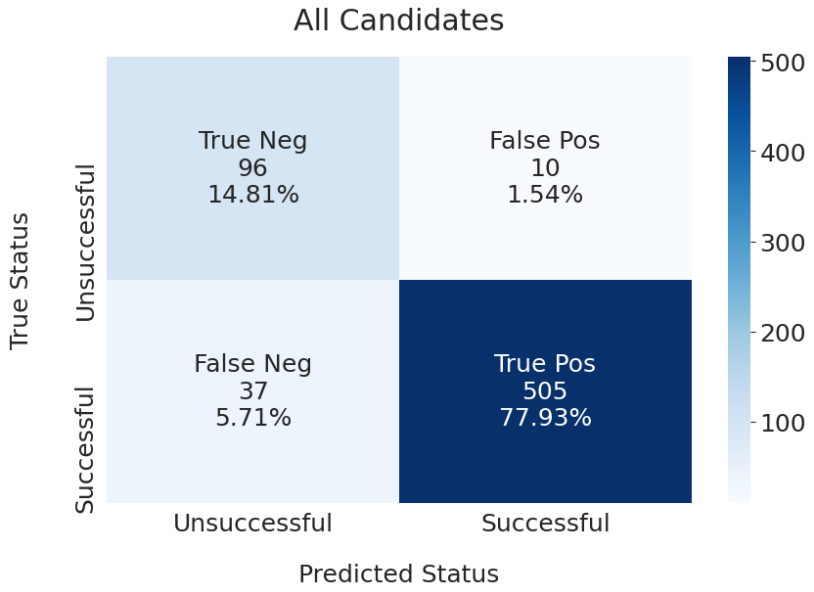

Figure 17. Confusion Matrix of Job Success prediction - XGBoost model

This experiment compared the effectiveness of a fully trained machine learning model with traditional manual recruitment methods.

➢ *Findings:*

- The human recruiters had an accuracy rate of *84%*. This means that they could correctly identify a candidate's potential success in a role *84%* of the time.
- The machine learning model, on the other hand, boasted an accuracy of *92.7%*, which is significantly higher.
- A key metric for businesses is the "False Positive" rate, which refers to candidates who are expected to do well but do not. The machine learning model had a lower rate of such candidates, making it more cost-effective in the long term.

3. Efficiency and speed

The effectiveness of machine learning was a major theme in this study.

❖ *Findings:*

- The machine learning model can process and predict candidate success in seconds.

- In contrast, manual recruiting procedures can take days, which is frequently a limitation due to the availability of human recruiters.

As a general conclusion, this case shows that with the integration of machine learning and natural language processing into the recruitment process, companies can better identify suitable candidates with higher accuracy, but they can also do so much faster than traditional manual methods.
Future paradigm shifts brought about by technology might redefine and enhance recruitment practices.

3.5.1 Challenges

- *Challenges in Data Collection:* Obtaining comprehensive data for a candidate often involves pulling from multiple sources. This multi-source gathering demands rigorous data processing because inaccuracies or inconsistencies can skew predictions and lead to suboptimal hiring decisions.
- *Machine Learning's Perceived Credibility:* While there's increasing interest in harnessing machine learning for recruitment, its trustworthiness is still debated. Many organizations use it as a supplementary tool, with human recruiters retaining the final say in candidate selection.
- *The Elusiveness of Defining Success:* Pinpointing what constitutes "job success" is complex, varying from one organization to the next. It is challenging to distinguish why a candidate is deemed unsuccessful. For instance, was it a shortfall in their skills or did they simply receive another job offer elsewhere? Both scenarios lead to the same label, "Unsuccessful," yet the underlying reasons are vastly different.

Thesis Conclusions

In this research, the primary objective is to assess the applicability and benefits of Machine Learning (ML) and Natural Language Processing (NLP) in the recruitment process, by providing a holistic overview of existing methods, pushing the boundaries of these techniques with novel optimization strategies.

The first approach was a combination of Word2vec and the SVM method, with the aim of creating a system that could systematically analyze a pool of candidates and match their qualifications with job titles and industry-specific details. The integration of these two algorithms brought a noticeable improvement in candidate-job matching, speeding up the process about 80% faster than traditional recruiting. This approach showed that even simple technological integration could yield substantial benefits.

Another method combination explored is BERT and XGBoost, with the goal of further elevating precision in identifying job title similarities. With its advanced capabilities, BERT delves deep into textual data, uncovering patterns and nuances that simpler models might overlook, making it an ideal choice for understanding diverse job descriptions and varied resumes, while XGBoost enhances the decision-making process. This integration managed to enhance the system's accuracy by an additional 2% while operating more efficiently and requiring less memory, thereby establishing it as a cost-effective solution for large-scale recruitment operations.

Beyond the initial sorting based on job titles, the essence of recruitment lies in identifying candidates' potential for success. Thus, the research introduced a comprehensive evaluation system. This new approach not only assessed candidates on their technical or specific skills but also evaluated their general qualities, such as teamwork, adaptability, and problem-solving. The holistic evaluation led to a 5% improvement in accuracy. Even more impressively, when compared to recruitment by humans, the system showed an accuracy rate of 92.7%, showing that the method is reliable and has a lot of potential.

Putting together all of these research findings and observations, it is clear that using machine learning and natural language processing in hiring is not just a concept for the future, but a real, useful way to solve problems in the present. This research not only presents an innovative perspective on recruitment but also offers actionable tools and methodologies that can streamline and enhance the talent acquisition process for organizations of all sizes.

Thesis Contributions

1. Analysis of the application of Machine Learning (ML) and Natural Language Processing (NLP) techniques in the recruitment industry

2. Suggested approach: Combination of Word2vec and the SVM method, with the aim of creating a system that could systematically analyze a pool of candidates and match their qualifications with job titles and industry-specific details.

3. Suggested approach: Combination of BERT and XGBoost method, with the goal of further elevating precision in identifying job title similarities.

4. Suggested holistic evaluation approach in recruitment for identifying candidates' potential for success

5. Formulated Job success prediction model

6. Created new perspective on how the recruitment industry should evolve to address the shifts in the globalization of the labor market across many professions, especially in the era of AI

List of publications in connection with the dissertation thesis

1. Mankolli E. M., Guliashki V. G. (2020), “Machine Learning and Natural Language Processing: Review of Models and Optimization Problems”, Proceedings of 12th ICT Innovations Conference 2020, held on 24-26 September 2020 in Skopje, Republic of North Macedonia, "Machine Learning and Applications", Vesna Dimitrova, Ivica Dimitrovski (editors), **Springer**, due to appear in Volume 1316 of the Communications in Computer and Information Science series (CCIS), **ISBN: 978-3-030-62097-4**, **ISSN: 1865-0937, SJR (0.188)**, Computer Science - **Quartile Q3,** pp. 71-86
   https://link.springer.com/chapter/10.1007/978-3-030-62098-1_7

2. Mankolli E., Guliashki V. (2021) “A Hybrid Machine Learning Method for Text Analysis to Determine Job Titles”, TELSIKS 2021, *Proceedings of papers of the “15th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services”*, October 20-22, 2021, Niš, Serbia, IEEE Catalog Number: CFP21488-USB, ISBN: 978-1-6654-2912-2 (IEEE), pp. 380-385, doi: 10.1109/TELSIKS52058.2021.9606341.
   https://ieeexplore.ieee.org/document/9606341

3. Mankolli E., Reducing the complexity of candidate selection using Natural Language Processing, In: *Proceedings of 29-th IEEE International Conference on Systems, Signals and Image Processing “IWSSIP 2022”*, June 01 - 03, 2022, Sofia, Bulgaria, pp. 1-4,
   doi: 10.1109/IWSSIP55020.2022.9854488.
   https://ieeexplore.ieee.org/document/9854488

4. Mankolli E., S. Bushati, Candidate Engagement Success Prediction Using Machine Learning and Natural Language Processing Techniques, In: *Proceedings of 24th Conference on Control Systems and Computer Science* (CSCS), May 24-26, 2023, Bucharest, Romania, pp. 431-435,
   doi: 10.1109/CSCS59211.2023.00074.
   https://ieeexplore.ieee.org/document/10214773

5. Guliashki V., E. Mankolli and S. Bushati, (2023), “A machine learning approach improving university campus security”, IEEE International Workshop on Technologies for Defense and Security TechDefense 2023, November 20-22, 2023, Rome, Italy, pp.341-345.
   https://www.techdefense.org/files/IEEETechDefense2023_FinalProgram.pdf