BRIDGING THE IT SKILL GAP WITH INDUSTRY DEMANDS: AN AI-DRIVEN TEXT MINING APPROACH TO JOB MARKET TRENDS USING LARGE LANGUAGE MODEL

TODSANAI CHUMWATANA[1], AYE KHIN KHIN HPONE [2]

[1]Department of Information and Communication Technology, International College, Rangsit University, Pathum Thani, Thailand

[2] Department of Information and Communication Technology, International College, Rangsit University, Pathum Thani, Thailand

E-mail: [1]Todsanai.c@rsu.ac.th, [2]Ayekhinkhinphone@gmail.com

ABSTRACT

The growing influence of technological advancements and the complexity of business requirements have revolutionized the IT industry underscoring the critical need for a comprehensive understanding of the skills and tools most valued and prioritized by employers. However, identifying these in-demand competencies remains a challenge given the dynamic nature of the job market. With previous studies exploring IT skill demands and very few studies utilizing AI-driven text-mining techniques to systematically analyze a large number of job advertisements, this research aims to address the ongoing gap between individual competencies and industry requirements. Leveraging text mining techniques, natural language processing (NLP), and large language models (LLMs) such as GPT-based AI model to enhance keyword extraction, this study provides the workflow process for generating knowledge-based insights using Term Frequency-Inverse Document Frequency (TF-IDF) scoring from a collection of IT job advertisements. The study identified vocabularies related to key skills and tools, by effectively summarizing the occurrences in job postings and revealing key terms that characterize specific roles and industry demands across IT sectors such as Software Development, DevOps, Cloud, Database Administration, Business Analysis, and Business Intelligence. Key findings reveal the demand for programming languages like VB.NET, Bash, and Python, alongside specialized languages including Dart, Kotlin, and Ruby, reflecting the growing importance of niche expertise. IT professionals are expected to be proficient in tools such as MS Visual Studio, and Crystal Report, as well as emerging frameworks like Flutter while skills in NoSQL and tools like MS Excel, ERP platforms, and Vertica are vital for supporting data-driven decision-making and business intelligence. By offering actionable insights into evolving industry expectations, this research serves as a valuable resource for educators with curriculum development, for recruiters with their talent acquisition strategies, and for job seekers with acquiring in-demand skills, ultimately ensuring a more responsive and competitive IT labor market.

Keywords: *Generative AI, Keyword Extraction, IT Employment, Large Language Models (LLMs), Natural Language Processing (NLP), Data Visualization, Data Analytics, Text Mining*

1. INTRODUCTION

The global information technology (IT) sector is experiencing unprecedented growth, driven by rapid advancements in technology and an increasing reliance on digital infrastructure. This growth has heightened the demand for skilled IT professionals who are proficient in a range of technical skills and tools. However, the swift pace of technological change is a recurring difficulty for instructors and educational institutions assisting students in adapting to this dynamic field [1]. A prevalent problem is the disparity between the curricula of these institutions and the demands of the business, hindering graduates' career prospects and employers' ability to locate suitably competent personnel [2],[3]. Likewise, this also poses a challenge for businesses, recruiters, and professionals attempting to keep pace with shifting skill demands. Traditional approaches for analyzing labor market needs, such as manual surveys, employer interviews, and industry reports, often suffer from small sample sizes, outdated data, and limited scalability. As a result, many hiring decisions and workforce planning strategies are based on incomplete or delayed information, leading to persistent skill gaps in the industry.

Job advertising, on the other hand, serves as a significant source of real-time data regarding employer requirements and expectations. This study uses text mining and natural language processing (NLP) approaches to analyze a diverse array of IT job posts, leveraging the advanced contextual understanding of large language models (LLMs) such as GPT-based AI model to parse large amounts of data, lessening the need for manual review and data extraction, particularly in labor-intensive tasks requiring multiple researchers [4], [5]. By utilizing Term Frequency-Inverse Document Frequency (TF-IDF) analysis, the study aims to quantify the relevance and demand of specific skills and tools across various IT job categories. The primary objective of this study is to identify and examine the skills and tools specified in IT job advertisements to determine the most sought-after qualifications according to current industry demands. This study seeks to bridge the gap by providing data-driven insights on the specific talents and tools most valued by employers, thereby aligning labor market supply and demand. The research evaluates the competencies prioritized by employers and aids in reconfiguring job roles and curricula to align sectors with graduates. By equipping students and graduates for the labor market by industry demands, this alignment enhances employability and fosters sustained economic growth through improved skills responsiveness from educational institutions [6]. Educators can leverage these insights to optimize courses following the changing demands of the world. This paper assists policymakers, recruiters, and job seekers by offering a clear representation of the evolving IT landscape, and illustrating methods to cultivate a more adaptive, responsive, and resilient labor market. The research eventually promotes sustained quality education and a workforce prepared for the contemporary economy.

2. RELATED WORKS

Several prior studies have been made in the area of aligning IT curricula with industry needs highlighting the gap between the skills imparted in educational institutions and those demanded by employers. For instance, Janicki et al. conducted a survey-based study indicating that academic curricula will need to shift towards contemporary IT skills to stay relevant in a field with rapid evolution where they found a rising popularity in emerging skills such as big data analytics and project management, with more traditional IT roles like software development progressing at a slower pace [7]. The advent of text analytics and data mining techniques has helped several researchers gain insight into job market trends, particularly within the IT sector. A study made by Maniu et al. [8] employed web crawling and text mining to analyze over 60,000 job postings in Romania and obtained frequently demanded skills such as teamwork and software. Similarly, Dong and Triche [9] applied text mining to more than 9000 job advertisements for data analyst positions, observing an increasing demand for skills such as Python and Tableau, while demand for skills such as SAP and Cognos has declined. This text-mining technique suggested that necessary skills required in the IT industry change over time and thus serve as a foundation for curriculum development aimed at bridging skill gaps.

In another study, Karakatsanis et al. [4] applied Latent Semantic Indexing (LSI) to match job advertisements with job descriptions in the O*NET database, giving industry-specific insights into the skills needed for different types of IT positions. Likewise, Lyon et al. [10] analyzed job postings from Indeed.com and found that employers generally indicate the need for graduates to have skills in data analysis and workplace readiness. Such studies emphasize the scope of the curriculum focusing on technical and professional competencies that should be imparted to the students, ensuring that graduates are well-prepared to meet employer expectations. Van-Duyet et al. [3] introduced a neural network model inspired by Word2Vec, called Skill2Vec, to map job skills into vector spaces, enhancing skill relevance matching in recruitment processes. More recently, Maree and Shehada [11] compared classical ML methods like logistic regression and decision trees with advanced large language models (LLMs) for resume-job matching. Their findings concluded the superior ability of LLMs in processing unstructured text which improved both accuracy and scalability in resume matching. In addition, Padmaja et al. [12] also applied NLP techniques like S-BERT for automated resume screening while Zhao et al. [13] introduced an embedding-based recommender system utilizing job-skill networks. These studies showcase how ML and NLP can enhance the hiring process by accurately mapping required job skills to candidate qualifications and offer scalability and precision, providing recruiters with a streamlined tool for skill-matching. One notable framework study by Almaleh et al. [14] proposed the Align My Curriculum (AMC) framework, which combines data scraping and feature engineering with Naive Bayes classification of topics and cosine similarity in an attempt for skill comparison to guide curricula set by educational institutions with market needs through a mechanism to update course content regularly.

Given the limitations of traditional job market analysis methods, often relying on static datasets, manual analysis, or qualitative insights which are ineffective for capturing real-time skill demand trends, the reviewed studies highlight the importance of analysis-driven techniques to effectively assess and map skills for capturing real-time demand trends. This paper extends the prior methodologies by utilizing TF-IDF analysis on job postings in the IT market, including the use of large language models (LLMs) and provides key insights for stakeholders to comprehend which skill is currently offered in large amounts and to be referenced to educational institutions, recruiters, and job seekers. The result from the study can be used as an indispensable guidance pointing out the gap between workforce capabilities and employer expectations, enabling stakeholders to make better informed decisions based on current industry demands and hiring patterns.

3. METHODOLOGY

The study employs a text-mining approach to analyze IT job advertisements based on skills and competencies. The proposed technique follows 7 main steps, consisting of (3.1) Data Source Identification, (3.2) Data Collection, (3.3) Data Cleaning and Preprocessing, (3.4) Data Preparation, (3.5) Data Modelling and Analysis, (3.6) Database, and (3.7) Data Visualization, as illustrated in Figure 1. This systematic approach helps the study to extract relevant information, categorize job roles, and identify in-demand skills for each IT role, ensuring a comprehensive analysis of the skills landscape in the IT sector.

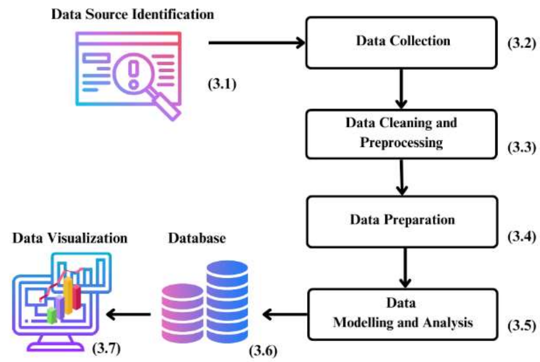


Figure 1: The architecture of the proposed technique

3.1. Data Source Identification

This research gathers data from some of the greatest job search websites, such as JobsDB, Glassdoor, and LinkedIn. JobsDB focuses on the Asia-Pacific region and gives a more local view of the IT job demands, allowing for better analysis of regional employment trends [15]. Glassdoor adds context to this data with its company reviews and salary reports, providing insights into further opportunities in the labor market, and employer expectations in different geographies [16]. In terms of providing key insights with its global coverage and professional network, LinkedIn serves as a valuable platform for an analysis that captures emerging trends in skill demands and employment patterns worldwide [17]. This multi-source approach is critical for better comprehensibility of the IT job market.

3.2. Data Collection

Data is collected by extracting job postings from several popular job platforms. A scripted browser is used for data extraction, developing a Python-based web crawler to extract job postings related to the Information Technology (IT) sector. It is programmed to navigate many job listings and retrieve details, such as job titles, job descriptions, skills needed, and tools indicated within a job posting, as depicted in Table 1 and Figure 2.

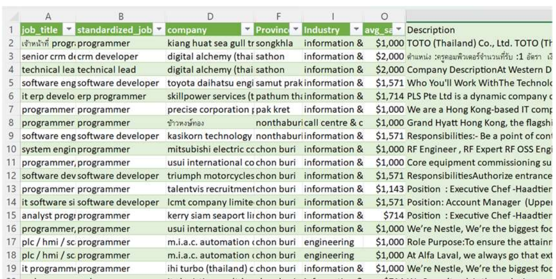
The data collection process extracted two months of job data where 3,167 job postings, in total, were gathered. This organized data collection procedure facilitates practical insights into current industry needs and elucidates the evolution of IT positions in response to demand.

Journal of Theoretical and Applied Information Technology
31st March 2025. Vol.103. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Extracted Job Data Attributes

Attribute	Description
Job Title	Specifies the advertised role (e.g., Software Engineer, Data Scientist).
Company	Identifies the hiring organization
Location	Indicates the job location, supporting regional analysis and insights
Job Description	Outline roles and responsibilities, providing context on required skills and tools.
Employment Details	Includes industry, job type (e.g., full-time, part-time), salary range, etc.



	A	B	D	F	I	O	
1	job_title	standardized_job	company	Provinc	Industry	avg_sa	Description
2	เจ้าหน้าที่ progr	programmer	kiang huat sea gull tr	songkhla	information &	$1,000	TOTO (Thailand) Co., Ltd. TOTO (Th
3	senior crm de	crm developer	digital alchemy (thai	sathon	information &	$2,000	ตำแหน่ง :ครูคอมพิวเตอร์จำนวนที่รับ :1 อัตรา
4	technical lea	technical lead	digital alchemy (thai	sathon	information &	$2,000	Company DescriptionAt Western D
5	software eng	software developer	toyota daihatsu engi	samut prak	information &	$1,571	Who You'll Work WithThe Technolo
6	it erp develo	erp programmer	skillpower services (t	pathum tha	information &	$1,714	PLS Pte Ltd is a dynamic company c
7	programmer	programmer	precise corporation p	pak kret	information &	$1,000	We are a Hong Kong-based IT comp
8	programmer	programmer	ข้าวหงษ์ทอง	nonthaburi	call centre & c	$1,000	Grand Hyatt Hong Kong, the flagshi
9	software eng	software developer	kasikorn technology	nonthaburi	information &	$1,571	Responsibilities:- Be a point of cont
10	system engin	programmer	mitsubishi electric cc	chon buri	information &	$1,000	RF Engineer , RF Expert RF OSS Engi
11	programmer,	programmer	usui international co	chon buri	information &	$1,000	Core equipment commissioning su
12	software dev	software developer	triumph motorcycles	chon buri	information &	$1,571	ResponsibilitiesAuthorize entrance
13	programmer	programmer	talentvis recruitment	chon buri	information &	$1,143	Position : Executive Chef -Haadtier
14	it software si	software developer	lcmt company limite	chon buri	information &	$1,571	Position: Account Manager (Upper
15	analyst progr	programmer	kerry siam seaport lir	chon buri	information &	$714	Position : Executive Chef -Haadtier
16	programmer,	programmer	usui international co	chon buri	information &	$1,000	We're Nestle, We're the biggest foo
17	plc / hmi / sc	programmer	m.i.a.c. automation (	chon buri	engineering	$1,000	Role Purpose:To ensure the attainr
18	plc / hmi / sc	programmer	m.i.a.c. automation (	chon buri	engineering	$1,000	At Alfa Laval, we always go that ext
19	it programm	programmer	ihi turbo (thailand) c	chon buri	information &	$1,000	We're Nestle, We're the biggest foo

Figure 2: Overview of extracted IT job data

3.3. Data Cleaning and Preprocessing

Job postings are cleaned and preprocessed by transforming the raw data, with noisy items, into clean data to be analyzed efficiently which includes converting text to lowercase, removing special characters and punctuation, eliminating common stop words and duplicate entries, correcting spelling errors, and filtering out non-English text to maintain a clean, uniform dataset ready for keyword extraction and analysis [18]. These processes were implemented using Python's libraries, such as Pandas for handling missing values and data imputation, along with NLTK (Natural Language Toolkit), Text Blob, and regex for text processing. Additional data preprocessing was performed to correct inconsistencies and errors, such as typographical and formatting issues. Location data was split into districts and provinces to facilitate location-based analysis and relative time indicators (e.g., "30d+" or "15d+") were converted into exact date formats, resulting in a timeline of each job posting, and non-English entries were filtered out applying language recognition. This step allows the dataset was optimally prepared for the next stages of analysis, ensuring the accuracy of the results.

3.4. Data Preparation

The data preparation phase was critical for structuring and organizing the extracted job data to enable a meaningful analysis. This process started with categorizing job titles based on predefined IT categories. These categories were developed using established frameworks from ABET (Accreditation Board of Engineering and Technology), AIS (Association for Information Systems), and ACM (Association for Computing Machinery), supplemented by professional judgment where necessary [19], [20], [21]. Subsequently, to indicate competencies and fields of expertise for each job title extracted from the dataset, guidance from the International Standard Classification of Occupations (ISCO-08) developed by the International Labor Organization [22] was followed. The main job categories for this study are Software Development, DevOps and Cloud, Database Administration, Business Analysis and Business Intelligence, Security, Network and System Administration, Architecture, and Help Desk and IT Support. These defined categories are further used to map job titles against skills and tools.

In the keyword extraction process, job descriptions were analyzed to find critical skills and tools linked to each job title within the designated IT categories. To further refine keyword extraction and synonym analysis, natural language processing (NLP) and large language models (LLMs) such as GPT-4o were employed. These models were accessed via API integration, with Python's requests library to submit prompts, retrieve responses, and extract keywords from job description texts, as in Figure 3. Carefully designed prompts directed the model to identify essential skills, tools, and competencies in job descriptions and to categorize related or synonymous terms. For example, prompts included, "List the primary skills, tools, and competencies described in the following job posting" and "Identify related or synonymous terms for each extracted skill.". Prompts were iteratively refined to enhance clarity and relevance, achieving a high level of accuracy in keyword extraction.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```python
import openai

openai.my_api_key = 'YOUR_API_KEY'

messages = [ {"role": "system", "content": "You are a keyword extractor."} ]

while True:
  message = input("User : ")
  if message:
    messages.append(
      {"role": "user", "content": message},
    )
    chat = openai.ChatCompletion.create(
      model="gpt-4o", messages=messages
    )

  reply = chat.choices[0].message.content
  print(f"ChatGPT: {reply}")
  messages.append({"role": "assistant", "content": reply})
```

Figure 3: Example of keyword extraction

Post-processing of model outputs involved parsing responses to organize and clean the extracted keywords. This step included removing duplicates, standardizing terminology, and grouping similar terms, which ensured consistency across the dataset. To address potential biases or misinterpretations by the large language models (LLMs), manual human verification was conducted, allowing for adjustments that maintained alignment with the study's objectives and minimized any bias in keyword extraction.

The use of LLMs facilitated the interpretation of complex textual information related to job titles and data sources, allowing for more accurate categorization and identification of essential skills and tools. By recognizing synonyms and grouping related phrases under a single skill category (e.g., grouping "data visualization," "data dashboards," and "Power BI" as "data visualization"), LLMs reduced redundancy and improved the accuracy of skill-demand analysis. A data dictionary was also developed during this phase to formalize the extraction process, ensuring consistency in categorizing skills and tools. This ensured that the approach was less prone to human error and bias, further validating data integrity.

3.5. Data Modelling and Analysis

Text mining is a subset of data mining aimed at extracting meaningful information from unstructured text data through various analytical methods [18]. The initial phase of the data analysis includes listing the terms and turning them into a word frequency format represented as a corpus, followed by the construction of a document-term frequency (DFM) matrix that catalogs text documents as word counts— a method known as the "bag-of-words" technique. A frequency table was created for each keyword in the dataset, illustrating its distribution across various jobs to determine initial term weights.

For analysis, Term Frequency-Inverse Document Frequency (TF-IDF) scoring was used to measure the weight of skills and tools within each IT job category [23]. TF-IDF scores were computed to weigh the terms found in job descriptions and identify the skills and tools that are more relevant for each job title. The TF-IDF score is calculated using the following equations:

$$TF\ (t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad (1)$$

$$IDF(t) = log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ term\ t}) \quad (2)$$

$$TF\text{–}IDF\ (t,d) = TF(t,d)\ \times\ IDF(t) \quad (3)$$

This scoring method ranks skills by relevance, with higher scores indicating terms that are both frequently mentioned within job descriptions for specific titles and relatively unique to those titles within the dataset.

3.6. Database

A database schema was designed to store the cleaned, categorized, and analyzed data, as illustrated in Figure 4. For efficient query and retrieval, Structured Query Language (SQL) was used to handle database operations, including Data Definition Language (DDL) statements for schema creation and Data Manipulation Language (DML) statements for data insertion, updating, and querying. Using the structured approach, and the assistance of database schema (Figure 4), this study effectively identifies essential skills and tools in the IT job market.

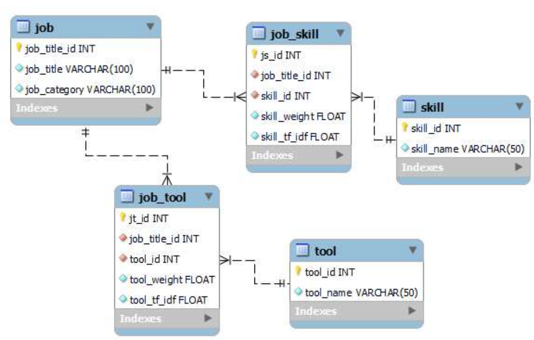


Figure 4: Database Schema

3.7. Data Visualization

The findings, visualized through data dashboards as in Figure 5 and Figure 6, reveal significant insights into the IT job market across diverse sectors. Visualization tools such as Power BI, Tableau, and Matplotlib, may be used to create intuitive and informative data reports to help stakeholders better interpret findings and uncover trends in the IT job market data.

Figure 5 and Figure 6 each show a comprehensive analysis of Job Demand, Skills, Tools, and Salary for *Software Development, DevOps, and Cloud Roles* and for *Database Administration, Business Analysis, and Business Intelligence Roles* respectively.

strategies as indicated by their substantial recruitment demands in the IT sector.

In the Software Development, DevOps, and Cloud sectors, Software Developers are the most sought-after professionals, comprising 39.45% of all job postings, followed by Programmers (16.02%), Cloud Engineer (7.81%) and Full Stack Developers (6.25%). This indicates that versatile and specialized professionals with competencies related to system design, programming, and cloud infrastructure are highly preferred.

Similarly, in the Database Administration, Business Analysis, and Business Intelligence sectors (Figure 6), Business Analysts dominate for more

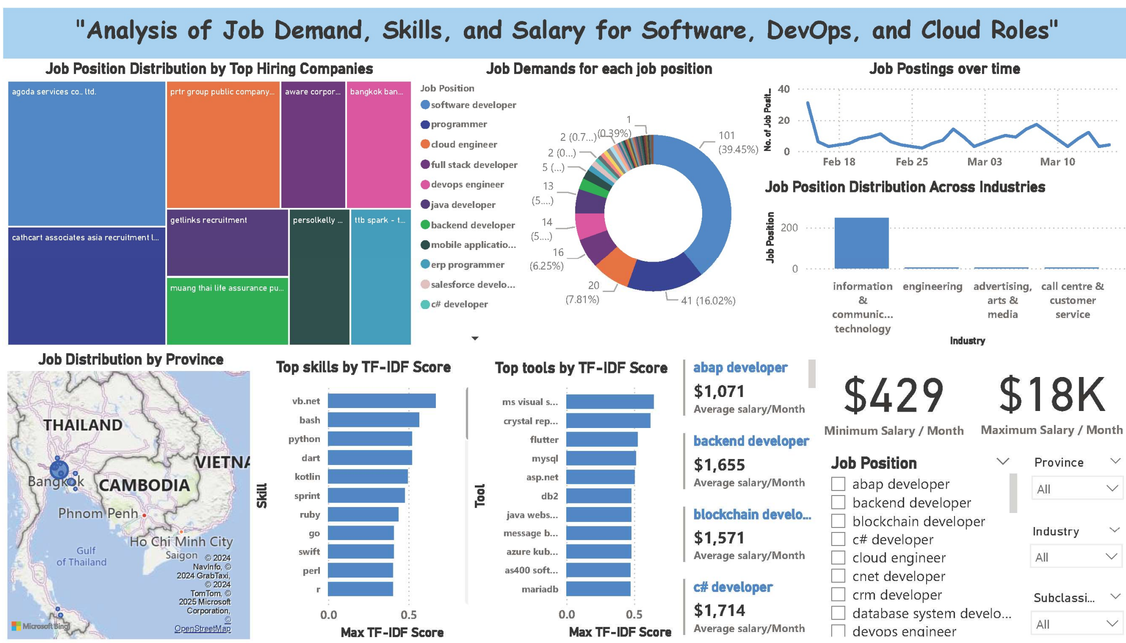


Figure 5: Analysis of Job Demand, Skills, Tools, and Salary for Software Development, DevOps, and Cloud Roles

Overview of Hiring Trends: Among all prominent hiring companies, Agoda Services Co., Ltd. and Cathcart Associates Asia Recruitment Ltd. emerge as the leading recruiter in Software Development, DevOps, and Cloud job roles as well as in Database Administration, Business Analysis, and Business Intelligence job roles. Other notable hiring companies for IT roles include PRTR Group Public Company Limited, and Aware Corporation Limited for the software sector roles, together with Bangkok Bank Public Company Limited, and Land and Houses Bank Public Company Limited making a strong showing in data-centric roles. This means that leading employers have focused recruitment

than half of the job postings by 52.98% with its critical role in bridging technology and strategic decision-making, followed by Data Analysts and Data Engineers, comprising 13.10% of postings. Furthermore, positions related to Data Science, Machine Learning, and Computer Vision have seen a relatively rising demand signaling the value placed on specialized expertise in the data-driven domain.

The distribution of job postings over time shows a fluctuating trend over several weeks, driven by project timelines, seasonal factors, and industry demands. These trends offer actionable insights for job seekers to align their applications with high-

Journal of Theoretical and Applied Information Technology
31st March 2025. Vol.103. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

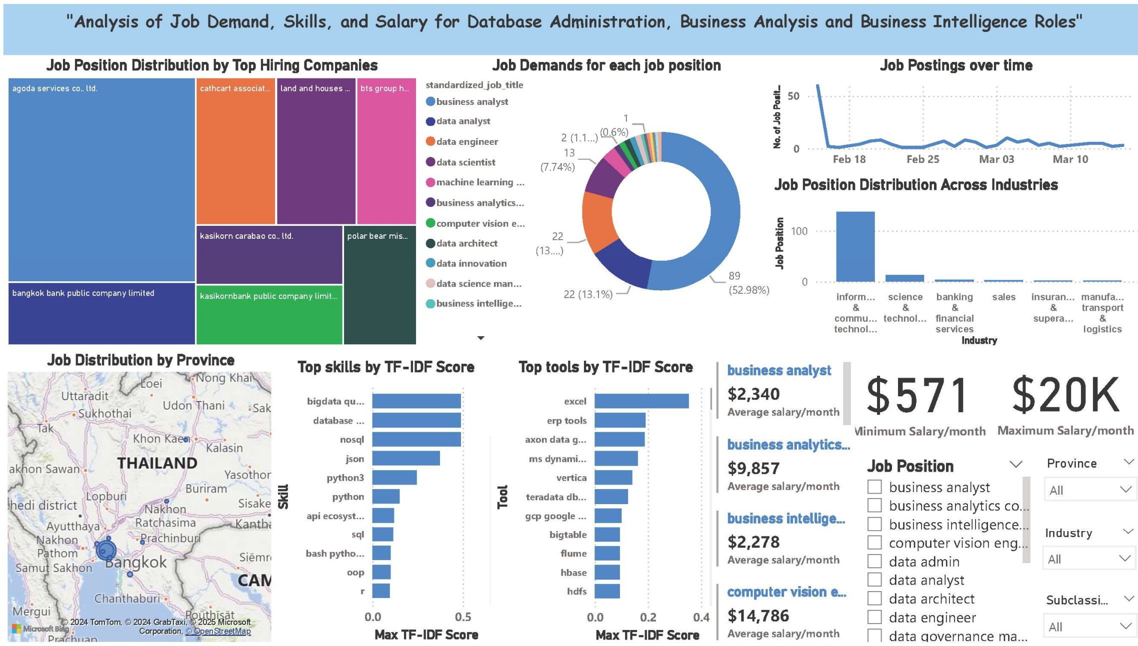


Figure 6: Analysis of Job Demand, Skills, Tools, and Salary for Database Administration, Business Analysis, and Business Intelligence Roles

demand periods and for employers to optimize recruitment strategies.

Skills and Tools in Demand: In the Software Development, DevOps, and Cloud job category, the demand for skills, quantified using TF-IDF values, reveals a strong preference for technical competencies such as VB.NET, Bash, and Python, reflecting their foundational role in Software and DevOps workflows. Kotlin, Dart, and Swift, which are essential for mobile application development, also rank highly, underscoring the importance of specialized language expertise in creating robust, user-friendly applications. Tools such as MS Visual Studio, Crystal Report, and Flutter underscore their significance in development workflows and cross-platform compatibility. Additionally, the high demand for advanced tools like Azure Kubernetes and MySQL highlights the industry's increasing reliance on cloud infrastructure and database management.

In Database Administration, Business Analysis, and Business Intelligence related job positions, skills such as Big Data Querying, Database Management, and NoSQL are critical for handling and analyzing large datasets. Programming skills in Python3, SQL, and API Ecosystem Development remain essential, while the demand for expertise in traditional tools like Excel and advanced platforms such as ERP Systems, Google Cloud Platform (GCP), and Teradata DB Systems underscores the importance of scalability and compliance in data analysis and governance.

Salary Analysis: Compensation trends vary across job categories and roles. The salary data analysis provides the compensation landscape ranging from a minimum salary of $429 (USD) to a maximum salary of $18,000 (USD) per month for the job positions in the Software Development, DevOps, and Cloud job category. Depending on individual expertise and specialization, ABAP Developers, for instance, can earn an average of $1,071(USD) per month, while Backend Developers can earn $1,654 (USD) per month. As expected with specialist or niche roles, Blockchain Developers, for instance, command reasonably high salaries, reflecting the priority level placed on specialized knowledge and experience.

In the Database Administration, Business Analysis, and Business Intelligence job category, salaries range from $571 (USD) to $20,000 (USD) per month. The most popular role, Business Analyst, earns an average of $2,340 (USD) per month, whereas specialized roles such as Business Analytics Consultant and Computer Vision Engineer command competitively higher salaries, highlighting their value to meet complex organizational challenges.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Geographic and Industry Distribution: Among all the job postings in Thailand, a geographic breakdown reveals that Bangkok, a well-known major city in the country, emerges as the central hub for IT opportunities. This signifies the concentration of IT infrastructure and industry within the region providing insights to professionals pursuing employment in high-opportunity regions and for policymakers striving to balance regional workforce development.

In addition to geographic trends, industry-specific distributions reveal the dominance of IT roles in the Information and Communication Technology industries across both categories, indicating the reliance of these technical roles' nature on innovation and operation. Other industries have comparatively fewer openings with Engineering, Advertising, Arts & Media, Call Center & Customer Service industries for the Software development roles and Science and Technology, Banking and Financial Services, and Sales industries for the analytics role.

Through data analysis and visualization, stakeholders can get valuable insights into current hiring trends, skill demands, and critical metrics such as geographic hiring patterns and salary benchmarks. As the IT industry continues to evolve, understanding dynamics is crucial for driving innovation and addressing workforce challenges. The visualizations presented in this section provide a high-level view of trends and demands in IT job roles, skills, and tools. The following section delves deeper into the experimental methodology and results that underpin these insights.

4. EXPERIMENTAL RESULTS

In this section, the experiment on IT job analysis based on the proposed technique is presented. The findings of the study revealed key competencies and technologies sought after in the IT job industry aimed to serve as a guidance on key industry requirements for job seekers and employers, as illustrated in Figure 7 and Figure 8. Term Frequency-Inverse Document Frequency (TF-IDF) is used to determine the relevance of terms where key terms with higher TF-IDF values are deemed more significant.

```python
import pandas as pd
import numpy as np
from google.colab import files
uploaded = files.upload() # Prompt the user to upload the Excel file
file_name = list(uploaded.keys())[0]  # Get the name of the uploaded file
df_raw = pd.read_excel(file_name, sheet_name=0)

skills = df_raw.columns[1:]
df_raw[skills] = df_raw[skills].apply(pd.to_numeric, errors='coerce').fillna(0)
df_raw['total_skills'] = df_raw[skills].sum(axis=1)

# Term Frequency (TF)
df_tf = df_raw.copy()
for skill in skills:
    df_tf[skill] = df_raw[skill] / df_raw['total_skills']
df_tf = df_tf.drop(columns='total_skills')

print("\n=== Term Frequency (TF) ===") # Preview TF
print(df_tf.head())

# Document Frequency (DF)
df_skill_appearances = (df_raw[skills] > 0).sum(axis=0)

# Inverse Document Frequency (IDF) using log base 10
N = len(df_raw)  # Total number of job titles
df_idf = pd.DataFrame({
    'Skill': skills,
    'DF': df_skill_appearances,
    'IDF': df_skill_appearances.apply(lambda x: np.log(N / (1 + x))) / np.log(N)
}).reset_index(drop=True)

print("\n=== Document Frequency (DF) ===")# Preview DF and IDF
print(df_idf[['Skill', 'DF']].head())
print("\n=== Inverse Document Frequency (IDF) ===")
print(df_idf[['Skill', 'IDF']].head())


df_tfidf = df_tf.copy() # Step 6: Calculate TF-IDF
for skill in skills:
    df_tfidf[skill] *= df_idf.loc[df_idf['Skill'] == skill, 'IDF'].values[0]

print("\n=== TF-IDF Matrix ===")# Preview TF-IDF
print(df_tfidf.head())

df_tf_only = df_tf.set_index('Job Title')
df_df_only = df_idf[['Skill', 'DF']].set_index('Skill')
```

Figure 7: Python code implementation for TF-IDF analysis

```
=== Sorted TF-IDF (Top 10) ===
                    Job Title                   Skills  TF-IDF Score
0              data architect                    nosql      0.488084
1   data management assistant      database management      0.488084
2  business analytic consultant          bigdata query      0.488084
3             data specialist                     json      0.372021
4             data specialist                  python3      0.244042
5   aws cloud and data engineer                 python      0.149726
6              data innovation  api ecosystem development   0.118863
7  business intelligence analyst                   sql      0.114531
8     computer vision engineer                     oop      0.099053
9     computer vision engineer                    bash      0.099053
```

Figure 8: Example output from TF-IDF analysis

Software Development, DevOps, and Cloud Roles

The analysis of Software Development, DevOps, and Cloud roles revealed the critical role of versatile programming languages and specialized tools in addressing the demands of modern workflows, as in Table 2. For instance, programmer roles show a TF-IDF score of 0.669 for VB.NET, emphasizing its significance in robust application development tasks. Similarly, Python emerges as a highly versatile language, with substantial TF-IDF scores for software developers (0.522) and programmers (0.433), underscoring its foundational role in building scalable and efficient systems. The work of a DevOps Engineer is closely tied to Bash (TF-IDF: 0.566) and Python (TF-IDF: 0.477), indicating an increasing demand for automation and scripting within DevOps practices.

The results further indicate a strong emphasis on Agile practices, as evidenced by Sprint's TF-IDF score of 0.477 for software developers. For Full Stack Developers, Ruby is a critical skill, achieving a TF-IDF score of 0.437. This highlights its role in creating robust back-end architectures and comprehensive full-stack solutions. Positions on mobile application development exhibit elevated TF-IDF scores of 0.522 for Dart, 0.494 for Kotlin, and 0.408 for Swift respectively. This signifies a robust need for language-specific competencies, highlighting companies prioritized technical proficiencies in this domain.

Table 2: Top Demanded Skills and TF-IDF Scores for Software Development, DevOps, and Cloud Roles

Software development, DevOps, and Cloud Roles			
	Standardized Job Tile	Skills	TF-IDF
1	programmer	vb.net	0.669
2	devops engineer	bash	0.566
3	software developer	python	0.522
4	mobile application developer	dart	0.522
5	mobile application developer	kotlin	0.494
6	devops engineer	python	0.477
7	software developer	sprint	0.477
8	full stack developer	ruby	0.437
9	programmer	python	0.433
10	mobile application developer	swift	0.408

The tools associated with these roles, as shown in Table 3, reinforce these findings. For programmers, tools like MS Visual Studio (TF-IDF: 0.646) and Crystal Report (TF-IDF: 0.620) dominate, reflecting their foundational importance in creating robust and efficient applications. Similarly, ASP.NET (TF-IDF: 0.503) and Java WebSphere (TF-IDF: 0.479) highlight the specialized environments preferred for enterprise-level development tasks. Mobile application developers demonstrate a strong reliance on Flutter (TF-IDF: 0.527), implying its value in cross-platform development. Backend developers and Java developers share a focus on Message Broker and DB2 (TF-IDF: 0.479 each), both pivotal tools for managing high-performance systems and databases. For software developers, MySQL (TF-IDF: 0.515) and integrated solutions like MS Power Automate (TF-IDF: 0.473) emphasize the need for versatile tools that streamline workflows and support integration. DevOps engineer's reliance on Azure Kubernetes (TF-IDF:0.478) underscores the growing importance of container orchestration and scalable cloud-based solutions in modern IT operations. This analysis reaffirms the role of tool diversity and specialization in meeting the specific requirements of these roles.

Table 3: Top Demanded Tools and TF-IDF Scores for Software Development, DevOps, and Cloud Roles

Software development, DevOps, and Cloud Roles			
	Standardized Job Tile	Tools	TF-IDF
1	programmer	ms visual studio	0.646
2	programmer	crystal report	0.620
3	mobile application developer	flutter	0.527
4	software developer	mysql	0.515
5	programmer	asp.net	0.503
6	backend developer	message broker	0.479
7	java developer	db2	0.479
8	programmer	java websphere (IBM WebSphere Application server)	0.479
9	devops engineer	ms azure kubernetes service	0.478
10	software developer	ms power automate	0.473

Database Administration, Business Analysis, and Business Intelligence Roles

The roles associated with data-centric domains, Database Administration, Business Analysis, and Business Intelligence, demonstrated a strong emphasis on advanced data management and analytical skills, as presented in Table 4. For roles like Data Architect and Data Management Assistant, the ability to work with advanced database systems is essential. Both positions emphasize expertise in NoSQL and Database Management, each achieving a TF-IDF score of 0.488, reflecting their importance in designing, managing, and optimizing modern database infrastructures. The role of a Business Analytic Consultant demands proficiency in Big Data Query (TF-IDF:0.488), underscoring the critical importance of expertise for managing large-scale data processing and therefore, supporting strategic business decisions effectively. The data specialist position exhibits the TF-IDF score of 0.372 for JSON and 0.244 for Python3, signifying that data organization and management are crucial components of this function.

Table 4: Top Demanded Skills and TF-IDF Scores for Database Administration, Business Analysis, and Business Intelligence Roles

Database Administration, Business Analysis, and Business Intelligence			
	Standardized Job Tile	Skills	TF-IDF
1	data architect	nosql	0.488
2	data management assistant	database management	0.488
3	business analytic consultant	big data query	0.488
4	data specialist	json	0.372
5	data specialist	python3	0.244
6	aws cloud and data engineer	python	0.150
7	data innovation	api ecosystem development	0.119
8	business intelligence analyst	sql	0.115
9	computer vision engineer	oop	0.099
10	computer vision engineer	bash	0.099

Additionally, the AWS Cloud and Data Engineer role emphasizes Python (TF-IDF: 0.150), further highlighting the importance of scripting and automation in cloud data engineering practices. The role of a Data Innovation Specialist prioritizes API Ecosystem Development (TF-IDF:0.119), underscoring the growing significance of integrating APIs for seamless data sharing and innovative applications. For a Business Intelligence Analyst, SQL (TF-IDF: 0.115) remains a fundamental skill, enabling the efficient querying and manipulation of databases to extract business-critical insights. In the specialized field of Computer Vision Engineering, the demand for Object-Oriented Programming (OOP) (TF-IDF: 0.099) and Bash (TF-IDF: 0.099) highlights the need for a robust understanding of programming paradigms and automation tools which are integral for building and optimizing computer vision systems that rely on scalable and efficient codebases.

Table 5: Top Demanded Tools and TF-IDF Scores for Database Administration, Business Analysis, and Business Intelligence Roles

Database Administration, Business Analysis, and Business Intelligence		
Standardized Job Tile	Tools	TF-IDF
data admin	ms excel	0.354
erp business process analyst	erp tools	0.191
data governance manager	axon data governance tool	0.188
erp business process analyst	ms dynamics 365	0.161
data science manager	vertica	0.140
technical lead	teradata db system premise	0.125
technical lead	gcp (google cloud platform)	0.101
data architect	bigtable	0.094
data architect	hdfs	0.094
data architect	hbase	0.094

For data-centric roles, table 5 presents the tools that are in high demand for Database Administration, Business Analysis, and Business Intelligence job categories. Higher TF-IDF scores signify the critical role these tools play in their respective job functions.

For Data Administrators, Microsoft Excel remains a key foundational tool (TF-IDF: 0.354) for handling day-to-day data administration tasks in managing, analyzing, and visualizing data. Individuals intending to work related to the Enterprise Resource Planning Field (ERP) must be able to use tools like the ERP systems (TF-IDF: 0.191) and MS Dynamics 365 (TF-IDF: 0.161) to ensure the seamless integration and management of enterprise processes for operational efficiency and alignment with organizational goals. Tools like the Axon Data Governance Tool (TF-IDF: 0.188) for data compliance and data quality are prominent for Data Governance positions. Likewise, the role of a Technical Lead demands expertise in advanced platforms such as Teradata DB System Premise (TF-IDF: 0.125) and GCP (Google Cloud Platform) (TF-IDF: 0.101) for handling powerful database systems and cloud platforms. For Data Science Managers, Vertica (TF-IDF: 0.140) stands out as a significant tool with its advanced analytics capabilities for deriving insights from large and complex datasets. Similarly, specialized tools such as Bigtable, HDFS, and HBase, each with a TF-IDF score of 0.094, associated with Data Architect roles are essential for managing large-scale, distributed databases. This reflects the need for expertise in tools designed to handle massive datasets in enterprise environments.

Table 6: Top Demanded Skills and Tools for Software Development, DevOps, and Cloud Roles

Software development, DevOps, and Cloud	
Top demanded skills	Top demanded tools
vb.net	ms visual studio
bash	crystal report
python	flutter
dart	mysql
kotlin	asp.net
python	message broker
sprint	db2
ruby	java websphere
python	azure kubernetes
swift	ms power automate

Table 7: Top Demanded Skills and Tools for Database Administration, Business Analysis, and Business Intelligence Roles

Database Administration, Business Analysis, and Business Intelligence		
	Top demanded skills	Top demanded tools
1	nosql	ms excel
2	database management	erp tools
3	bigdata query	axon data governance tool
4	json	ms dynamics 365
5	python3	vertica
6	python	teradata db system premise
7	api ecosystem development	gcp (google cloud platform)
8	sql	bigtable
9	oop	hdfs
10	bash	hbase

The findings from the analysis of IT job postings for different IT job categories can be summarized into top demanded skills and tools, as outlined in Table 6 and Table 7, revealing several interconnected trends shaping the evolving IT job market.

Key observations of the study point out the high demand for programming proficiency, such as VB.NET, Bash, and Python programming languages along with specialized languages like Dart, Kotlin, and Ruby programming languages, reflect the rising prominence of niche expertise. In terms of technical tool competencies, IT professionals must be familiar with a broad array of tools, including MS Visual Studio, Crystal Report, as well as emerging frameworks like Flutter. In addition to the need to stay competitive in the ever-evolving IT landscape, data-centric skills, including NOSQL, database management and Big Data Querying heavily rely on the areas of data-driven decision-making along with tools like MS Excel, ERP platforms, and Axon Data Governance tool for data governance and operational efficiency in supporting business intelligence and analytical functions.

This study presents a scalable, AI-driven approach to analyzing IT job market trends but has some limitations. The dataset, sourced from job postings on popular hiring platforms may not fully capture informal hiring or unadvertised roles. Besides, tracking job postings over a longer period of time would provide deeper insights into shifting skill demands. Although the research is limited by its reliance on job advertisements and geographic constraints, it effectively bridges the skills gap with industry demands avoiding any employability hindering occurrences that may otherwise have been caused when workers are not competent enough and adaptable to the evolving job landscape.

Unlike prior research, this study leverages text mining using large language model with GPT-based AI model for keyword extraction ensuring contextual understanding when defining skill relevance, offering a quantitative and structured workforce analysis. Thus, it helps significantly in the identification of emerging technologies that are often overlooked in traditional studies which usually focused on established programming languages.

It is imperative to bear in mind the evolving nature of IT skill demand and that it requires an ongoing analysis, and thus, future research should incorporate time-series tracking to monitor changes over time. Addressing further research gaps will enhance IT workforce planning and ensure better labor market alignment. In light of these findings, stakeholders of the IT job market can have a clear picture of the requisite skill sets currently valued in the market. It is high time professionals continuously adapt to evolving technologies and tools to remain relevant and effective within the IT sector.

5. CONCLUSION

The study contributes significantly along many dimensions through the analysis of the industry's changing requirements for a better understanding of the specific competencies valued, allowing for academic alignment to upskill and reskill for IT jobs [24], [25]. Beyond presenting findings, this study provides practical takeaways for educators, recruiters, and policymakers. Educational institutions can use these findings to align their training programs that closely mirror industry requirements and incorporate prioritized technical and software skills. This means that underrepresented or overlooked essential skills in academic programs are also identified, fostering an updated curriculum development that promotes the development of a diverse skill set, and thus offering up-to-date recommendations for instructors and academic institutions to drop outdated and irrelevant courses. Concerns regarding the rise of Artificial Intelligence and the threats posed by job displacement are directly addressed, ensuring employability and adaptability in the face of technological change by empowering learners with the requisite competencies to thrive in work environments increasingly dominated by AI. Besides, the study contributes to the development of curricula and training programs aligned with the Sustainable Development Goals (SDGs) established by the United Nations, specifically SDG 4 (Quality Education) and SDG 8 (Decent Work and Economic Growth), promoting long-term quality education [6]. Companies and recruiters, on the other hand, can use the standardized key terms (i.e. skills and tools findings) to tailor job descriptions that may otherwise not provide a representation of the exact skills and experiences looking for and improve recruitment strategies, allowing them to attract candidates with the necessary skill sets. By offering a systematic approach to aligning education with market expectations, this study contributes to better workforce planning, reduced skill mismatches, and improved economic sustainability.

This research significantly advances existing knowledge by introducing an AI-powered work flow process, overcoming the limitations of small-scale, and manually curated studies. The use of natural language processing (NLP), and large language models (LLMs) for job market analysis presents a reliable contribution to the field of labor economics, education policy, and AI-driven workforce analytics.

In the future, studies could expand on this work by uncovering emerging trends in specialized IT fields. It would also be valuable to investigate the impact of incorporating industry-aligned curricula on student outcomes, such as employability and job performance. Additionally, expanding the geographic and sectoral scope of the research could provide more generalized insights and strengthen the foundation for sustainable education reform. Exploring the role of AI and advanced analytics in further refining job market analysis would also present a promising avenue for future work, strengthening the foundation for long-term sustained quality education in the IT domain.

REFERENCES:

[1] M. Pejic Bach, T. Bertoncel, M. Meško and Ž. Krstić, "Text mining of industry 4.0 job advertisements," International Journal of Information Management, vol. 50, 2019.

[2] K. Gallagher et al., "A Typology of Requisite Skills for Information Technology Professionals," in 44th Hawaii International Conference on System Sciences (HICSS), 2011.

[3] L. Van-Duyet, V. M. Quan and D. An, "Skill2vec: Machine learning approach for determining the relevant skills from job description," arXiv preprint arXiv:1707.09751, 2017.

[4] I. Karakatsanis et al., "Data mining approach to monitoring the requirements of the job market: A case study," Information Systems, vol. 65, p. 1–6, 2017.

[5] OpenAI, "Gpt-4o overview," 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/. [Accessed May 2024].

[6] I. Makarenko, A. Plastun, Y. Petrushenko, A. Vorontsova and S. Alwasiak, "SDG 4 and SDG 8 in the knowledge economy: A meta-analysis in the context of post-COVID-19 recovery," Knowledge and Performance Management, vol. 5, no. 1, p. 50, 2021.

[7] T. Janicki, J. Cummings and D. Kline, "Information Technology Job Skill Needs and Implications for Information Technology Course Content," AITP-EDSIG, vol. 12, pp. 59-70, 2014.

[8] I. Maniu, E. L. Pop, A. Ratiu and E. T. Stefanescu, "Insights from IT jobs market with

text mining approach," SEA: Practical Application of Science, vol. 8, no. 3, 2020.

[9] T. Dong and J. Triche, "A longitudinal analysis of job skills for entry-level data analysts," Journal of Information Systems Education, vol. 31, no. 4, p. 312, 2020.

[10] L. Lyon, E. Mattern, A. Acker and A. Langmead, "Applying translational principles to data science curriculum development," iPres 2015 Proceedings, 2015.

[11] M. Maree and W. A. Shehada, "Optimizing Curriculum Vitae Concordance: A Comparative Examination of Classical Machine Learning Algorithms and Large Language Model Architectures," AI, vol. 5, no. 3, p. 1377–1390, 2024.

[12] D. L. Padmaja, . C. Vishnuvardhan, G. Rajeev and. K. N. S. Kumar, "Automated resume screening using natural language processing," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 10, no. 3, 2023.

[13] J. Zhao et al., "Embedding-based recommender system for job to candidate matching on scale," arXiv preprint arXiv:2107.00221, 2021.

[14] A. Almaleh, M. A. Aslam, K. Saeedi and N. R. Aljohani, "Align my curriculum: A framework to bridge the gap between acquired university curriculum and required market skills," Sustainability, vol. 11, no. 9, p. 2607, 2019.

[15] JobsDB, "Jobsdb - Search jobs across Hong Kong, Singapore & Thailand," 2024. [Online]. Available: https://www.jobsdb.com/. [Accessed 20 March 2024].

[16] Glassdoor LLC, "Glassdoor | Job Search & Career Community," 2024. [Online]. Available: https://www.glassdoor.com. [Accessed 20 March 2024].

[17] LinkedIn, "LinkedIn - Official Site," 2024. [Online]. Available: https://www.linkedin.com. [Accessed 20 March 2024].

[18] C. C. Aggarwal and C. Zhai, Mining Text Data, New York, NY, USA: Springer, 2012.

[19] Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and IEEE Computer Society, "Computer science curricula 2013: Curriculum guidelines for undergraduate degree programs in computer science," Association for Computing Machinery, New York, NY, USA, 2013.

[20] The Joint ACM/AIS IS2020 Task Force, P. Leidig and H. Salmela, "A competency model for undergraduate programs in information systems," Association for Computing Machinery, New York, NY, USA, 2021.

[21] Task Group on Information Technology Curricula, "Information technology curricula 2017: Curriculum guidelines for baccalaureate degree programs in information technology," Association for Computing Machinery, New York, NY, USA, 2017.

[22] International Labour Organization, "International Standard Classification of Occupations (ISCO)," 2024. [Online]. Available: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/. [Accessed 10 May 2024].

[23] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning, 2003.

[24] V. Granville, Developing Analytic Talent: Becoming a Data Scientist, John Wiley & Sons, 2014.

[25] A. Amado, P. Cortez, P. Rita and S. Moro, "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis," European Research on Management and Business Economics, vol. 24, no. 1, p. 1–7, 2018.