ICDSOS
The 3rd International Conference
on Data Science and Official Statistics
November 27 - 28, 2025

"From Data to Impact: Leveraging Innovation in Data Science and Official Statistics for Data-Driven Policy Making and Global Competitiveness"

Job Competency Extraction in Information and Technology Sector Using K-Means and Non-Negative Matrix Factorization (NMF) Algorithms

Alfitra R Geandra[1], Amir M Siregar[1] and Rani Nooraeni[1,*]

[1] Politeknik Statistika STIS, Jakarta, Indonesia

*Corresponding author's email: raninoor@stis.ac.id

Abstract. The advancement of information technology has led to a surge in online job vacancy data, which contains valuable information about the skill demands in the digital labor market. This study aims to extract job competency in the information and technology sector using a combination of K-Means clustering and Non-Negative Matrix Factorization (NMF). A total of 350 job postings were collected from the Kalibrr platform and processed through web scraping, text preprocessing, and feature representation using TF-IDF. The clustering results indicate that the optimal configuration consists of 10 clusters, as evaluated using the Silhouette Score and Davies-Bouldin Index. Each cluster represents a specific job topic, such as backend development, data science, QA automation, cybersecurity, and digital marketing. The results offer a structured overview of digital skill demands and can be utilized by educational institutions, training providers, and labor policy makers. However, the dataset's limited size, reliance on a single job platform, and the use of traditional machine learning techniques may not capture all semantic variations and complexities present in the broader job market. Consequently, future work should involve larger and more diverse datasets as well as advanced deep learning text representation approaches to enhance the robustness and generalizability of the results.

Keyword: competency extraction, digital labor market, K-Means, NMF, text mining.

1. Introduction

The advancement of information and communication technology (ICT) has revolutionized various aspects of human life, including the dynamics of the labor market. ICT enhances workforce productivity, expands employment opportunities, and shifts skill demand, while also triggering job polarization and requiring adaptations in both policy and education [1]. Digital transformation has driven a shift in recruitment processes from conventional print-based approaches to online platforms, such as websites and social media [2]. This digitalization has resulted in a large volume of job vacancy data, which not only reflects real-time labor market needs but also contains valuable insights into skill demand [3]. However, the utilization of such data remains limited among job seekers, employers, and policymakers. The increasing volume of job vacancy data, without a structured classification system, has introduced new challenges. Job seekers often struggle to filter information relevant to their competencies, while employers face difficulties in identifying the most suitable candidates from thousands of applicants. This information asymmetry is a major contributor to the phenomenon of skill mismatch—the misalignment between the skills possessed by the workforce and the actual needs of the labor market. This issue is particularly critical in Indonesia's human resource development, especially in the rapidly growing information and technology sector, which is still facing a shortage of skilled talent.

A report by the World Bank indicates that more than half of Indonesian workers experience a mismatch between their education, skills, and job requirements [4]. Meanwhile, the Ministry of Information and Communication Technology estimates that Indonesia faces a shortage of approximately 600,000 digital talents each year [5]. This imbalance underscores the urgency of a data-driven and adaptive labor market information system to effectively map skill requirements by sector and region. Although online job advertisements are granular and dynamic in nature, their use in academic research and labor policy formulation remains limited. Lukauskas et al. emphasized the importance of systematically clustering job vacancies based on core skills to generate structured, accurate, and actionable information [3]. This information is not only vital for policymakers but also for educational and training institutions in aligning curricula with current industry needs. Graetz, Restrepo, and Skans highlighted that modern technologies such as machine learning and robotics are accelerating changes in job structures and skill demands [6]. They showed how technology is reshaping labor markets through sectoral shifts, changing job types, and systematically redefining the skills required by employers. Therefore, understanding how technology reshapes the labor market is essential for designing relevant employment policies. Furthermore, Kobayashi noted that job advertisements are not merely recruitment tools, but also valuable data sources for occupational analysis [7]. Using machine learning-based classification models, information regarding job duties, required skills, and candidate attributes can be extracted automatically from job posting data. This approach enables the efficient creation of skill taxonomies and job clustering, in contrast to traditional job analyses, which are time-consuming and often become obsolete.

To address these challenges, unsupervised learning has emerged as a relevant approach for exploring hidden patterns in job vacancy text data. One widely used algorithm is K-Means, which clusters data based on feature similarity. In addition, Non-Negative Matrix Factorization (NMF), a topic modeling technique, can be used to identify dominant skills emerging within each job group, thereby adding a semantic dimension to the clustering results [8]. Several previous studies have demonstrated the potential of clustering in job advertisement analysis. Debao et al. used K-Means to cluster big data job vacancies in China and identified 10 primary clusters based on job descriptions [9]. Siswaja and Tri Prasetio also employed K-Means to map the spatial distribution of IT job vacancies and analyze concentration patterns by location [10]. Agustyani and Santoso applied Hierarchical Agglomerative Clustering to examine job advertisement characteristics on Jobstreet, focusing on educational background [11]. On the other hand, Kumar and Priya emphasized the importance of location and experience levels in clustering job postings and recommended the development of more targeted training programs [12]. Additionally, Lukauskas et al. combined NLP techniques with HDBSCAN to automatically generate real-time job profiles and demonstrated the potential of technologies such as BERT and UMAP for feature extraction [3].

Despite this progress, a notable gap remains in research that integrates clustering methods with topic modeling using NMF to identify dominant competencies within digital job clusters in a contextualized manner. Moreover, most prior studies have focused on spatial factors, educational backgrounds, or general job categories, rather than specifically addressing technical skills in Indonesia's digital sector. This study aims to segment job vacancies in the field of information and technology using a combination of K-Means and NMF algorithms. Data was collected from Kalibrr (https://www.kalibrr.com/id-ID/home), a leading digital recruitment platform that hosts thousands of job postings. The job filters applied included Front-End Developer, Back-End Developer, Data Scientist, Data Analyst, and Researcher to ensure data relevance to the research focus. The dataset consists of semi-structured information such as job titles, task descriptions, and qualification requirements. Through this approach, the study is expected to produce a systematic and informative mapping of the skills demanded by the digital industry. The results are anticipated to contribute to the development of labor market information systems and assist educational and training institutions in designing curricula that are more responsive to industry needs.

2. Research Method

This study adopts a quantitative descriptive approach aimed at exploring and analyzing thematic patterns in online job postings, particularly in the fields of information technology and data science. The main methods employed include web scraping to collect job vacancy data, text preprocessing to ensure the data is ready for further analysis, clustering to group job postings based on textual similarity, and topic modeling to identify the main topics within each job cluster. Figure 1 illustrates the research workflow used in this study.

2.1 Data collection

The data used in this study were obtained from the online job portal Kalibrr, accessed through https://www.kalibrr.com/id-ID/home. Data scraping was conducted during the period 25 June to 30 June 2025, ensuring that the dataset reflected job postings available within that timeframe. Filters for Front End and Back End Developer as well as Data Science, Analyst, and Researcher were applied to ensure that the collected job postings aligned with the research design. A total of 350 job postings were successfully collected. The collected data were semi-structured, containing information on job titles, company, work descriptions, and qualifications. Data collection was carried out using Python-based web scraping techniques, employing Selenium and Undetected Chromedriver libraries to navigate web pages, BeautifulSoup4 to extract information from the DOM (HTML elements), and Pandas to store the job data. Scraping was conducted across multiple pages to gather a representative number of job postings. The data were then saved in .csv format for further analysis.

Table 1. Features in the job postings dataset.

Feature	Description
Job title	The name of the position or role being offered (e.g., "Software Engineer").
Company	The name of the company offering the job (e.g., "Traveloka", "Danareksa").
Work descriptions	A brief description of the job role and responsibilities.
Qualifications	A list of skills needed for the job, such as programming languages, data analysis, etc.

2.2 Method of analysis

The semi-structured data obtained from web scraping must first be transformed into a more structured format before advanced analysis can be conducted. This transformation process is necessary because job posting data often contains inconsistencies, unstandardized text, and irrelevant information that may reduce the accuracy of clustering and topic modeling results. To address this, the analysis in this study was carried out through several sequential stages designed to convert the raw semi-structured dataset into structured analytical features, as described below.

2.3.1 Data preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis, as job postings tend to be unstructured, contain irrelevant words, and exhibit inconsistent formatting. The objective of this stage is to ensure that the textual data used for clustering and topic modeling is clean, consistent, and semantically meaningful [13]. The preprocessing process begins with language detection, as although most job descriptions on the Kalibrr website are written in English, there is still a possibility of encountering postings in Indonesian. Language detection is conducted using the langdetect library. Any postings identified as being in Indonesian are subsequently translated into English.

Text normalization is then performed to standardize technical terms within the job postings, such as "PostgreSQL," "Java," and others. This is followed by text cleaning, which includes converting all text to lowercase, removing special characters such as punctuation and symbols, and eliminating stopwords—words that do not carry significant meaning. The cleaned text is then tokenized into unigram and bigram units. Finally, lemmatization is applied to reduce words to their base or dictionary forms (lemmas) while preserving their contextual meaning and grammatical roles, using WordNet Lemmatizer from NLTK library.

2.3.2 Text vectorization

After the data preprocessing stage, text vectorization is performed to convert textual information into a numerical representation suitable for use in machine learning algorithms. This step is essential in text analysis, as models cannot interpret raw text and therefore require numerical encoding [14]. In this study, the Term Frequency–Inverse Document Frequency (TF-IDF) method is employed.

TF-IDF is a widely used traditional method in text analysis due to its low computational cost, making it suitable for handling large-scale textual datasets. The output of TF-IDF can be directly used as input for clustering algorithms such as K-Means, as well as serve as the foundation for topic modeling techniques such as Non-Negative Matrix Factorization (NMF). The core concept of TF-IDF is to calculate the importance score of a term within a specific document relative to its frequency across the entire document corpus [15]. The formula used is as follows.

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right)$$

explanation:

$w_{i,j}$: TF-IDF score of the *i*-th term in the *j*-th document
$tf_{i,j}$: term frequency of the *i*-th term in the *j*-th document
N : total number of documents
df_i : number of documents that contain the *i*-th term

2.3.3 Clustering

The clustering method is used to group job vacancies based on the similarity of content found in their job descriptions and minimum qualification requirements. The clustering technique selected in this study is K-Means Clustering, one of the most commonly used unsupervised algorithms for feature-based segmentation tasks.

K-Means operates by partitioning the data into k groups based on the proximity of each data point to the cluster center (centroid), which is updated iteratively until convergence. Each iteration consists of two main steps: assigning data points to the nearest cluster using Euclidean distance, and recalculating the centroid as the average vector of all points within the cluster [16]. The data representation using the TF-IDF method serves as the input for the K-Means algorithm. The choice of K-Means is based on its efficiency in handling high-dimensional data and its ability to produce segmentation results that are intuitively interpretable [17].

One of the main challenges of K-Means is determining the number of clusters (k) in advance. To address this, auxiliary approaches such as elbow method, silhouette score, and Davies-Bouldin Index (DBI) are utilized. The elbow method was used as an initial exploratory tool to identify the optimal number of clusters by observing the inflection point in the within-cluster sum of squares (SSE) curve [18]. However, because the elbow method provides only a visual and heuristic indication, it was not used as the sole determinant. The final selection of the optimal number of clusters was based on quantitative validation using the

Silhouette Score and DBI, which offer more objective evaluations of cluster cohesion and separation [19], [20].

2.3.4 Topic modeling

Topic modeling is an exploratory technique in text analysis that aims to uncover latent structures or hidden topic patterns within a collection of documents. This stage is conducted to provide a thematic overview of each cluster. The results indicate the types of competencies or IT-related job fields that frequently appear in the analyzed job postings.

This study employs the Non-negative Matrix Factorization (NMF) method. NMF was chosen due to its ability to produce coherent and interpretable topic representations and its widespread use in prior research, such as by Egger & Yu [21], Gallego et al. [22], and Luo et al. [23]. The number of topics is determined using the coherence score metric, which evaluates the interpretability and semantic coherence among the words within a topic. The higher the coherence score, the better the quality of the extracted topics [24].

In several clusters, the coherence score was found to be negative, indicating low semantic similarity among keywords. In some cases, coherence scores may be negative especially when document subsets are small or vocabulary overlap is low [25]. This makes automated metrics unstable. Therefore, in this study, clusters with negative coherence were set to a default of three topics to maintain interpretability and avoid overfitting to noise.

The dominant keywords of each topic were extracted and then assigned descriptive labels to represent their thematic content. ChatGPT was used to suggest relevant job positions based on the dominant keywords, and the final labels were reviewed and refined by researchers to ensure context appropriateness. The use of ChatGPT was motivated by its efficiency in generating consistent annotations and reducing individual bias, while still maintaining human oversight in validating the results [26], [27].

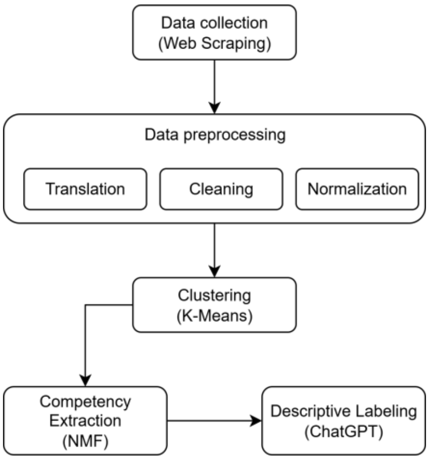
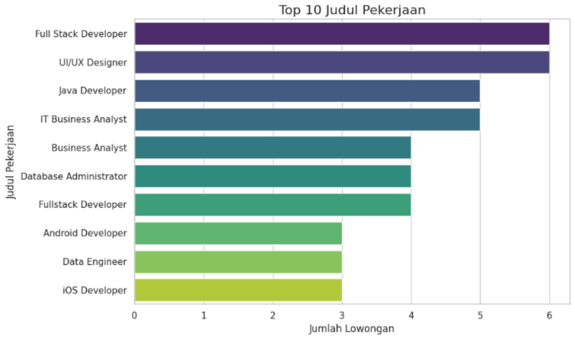


Figure 1. Research design.

3. Result and discussion

3.1. Job vacancy data exploration

A total of 350 job vacancy records were successfully collected. Each record includes information such as the job title, company name, job description, and minimum qualifications required. These postings originate from 82 unique companies, indicating a relatively diverse coverage across various industry sectors. Examples of data that have been collected can be seen in table 2.



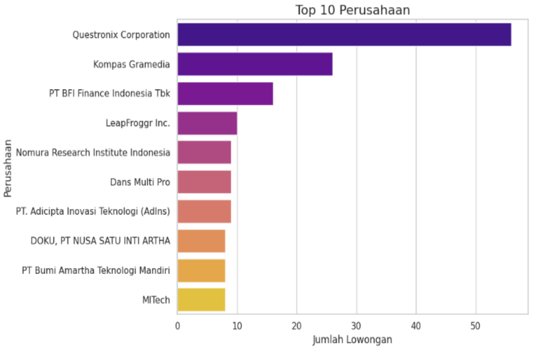
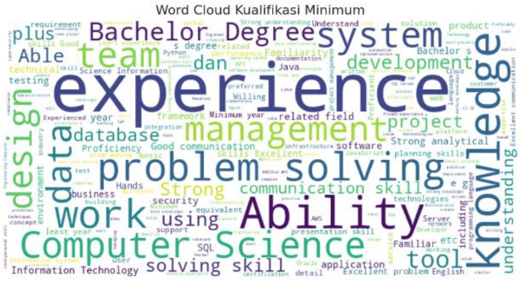


Figure 2. Job titles with the highest frequency.

Figure 3. Companies with the highest number of job postings.

Figure 2 shows that UI/UX Designer and Full Stack Developer are the two most sought-after positions, each with 6 job postings. Other popular positions include Java Developer and IT Business Analyst, indicating a high demand for professionals in digital product development and systems analysis. As shown in Figure 3, Questronix Corporation is the company with the highest number of job postings, totaling 56 vacancies, significantly surpassing other companies. Kompas Gramedia ranks second with 26 postings, followed by PT BFI Finance Indonesia Tbk with 16 postings.



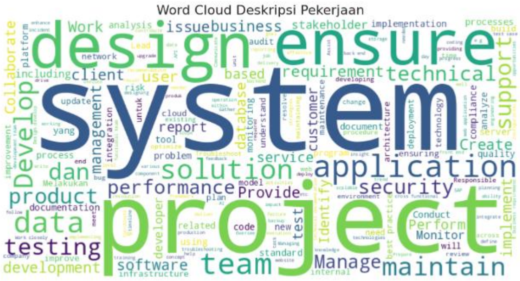


Figure 4. Word cloud for work descriptions.

Figure 5. Word cloud for qualifications.

Based on Figure 4, it can be concluded that the majority of job vacancies emphasize activities such as system design, project management, and the development of applications or solutions. Keywords such as “system”, “project”, “design”, and “ensure” dominate, indicating that companies are primarily focused on the development of complex information technology systems and the importance of thorough planning and project execution. Meanwhile, Figure 5 reveals that work experience and educational background in Computer Science are among the primary requirements. Terms like “problem-solving”, “management”, and “knowledge” highlight that applicants are expected to possess strong analytical abilities and technical proficiency. Other prominent terms such as Bachelor’s Degree, team, and communication skills also appear frequently, suggesting that formal education, collaborative ability, and effective communication are critical qualities sought by employers.

Table 2. Data collected examples.

Job Title	Company	Work Description	Qualification

IT Security Specialist	NTT Indonesia Technology	Perform vulnerability assessment and penetration testing on the defined scope, and provide a comprehensive the report. Perform security monitoring, data/log and forensic analysis, to proactively detect security incidents and threats. Plan for and perform periodic security reviews to validate that the security posture satisfies Information Security and facility security requirements	Bachelor's degree or equivalent in Information Technology or Computer Science. Minimum of 1 year work experience as security specialist. Experience in risk, compliance and information security policy development. Familiarity with security tools such as SIEM, antivirus, and firewalls. Ability to work in a fast-paced environment and adapt to changing priorities. Excellent problem-solving and communication skills.
Java Technical Lead	PGI Data	Design, develop, and implement efficient and scalable Java-based applications Collaborate with cross-functional teams to understand business requirements and translate them into technical solutions Write clean, well-documented, and maintainable code adhering to best practices Participate in code reviews and provide feedback to improve codebase quality	Diploma or Bachelor's degree in computer programming, computer science Min. experience 3 years as Tech Lead Java Experience handle Microservices, Kafka & RabbitMQ Deep knowledge of Postgre, Mysql and SQL Server Familiar with Java Spring Boot Familiar with Containerization (docker) Deep knowledge of tcp socket, web socket and messaging (Rabbitmq, kafka, nats) Familiar with google protocol buffer Familiar with TDD Deep knowledge of keycloak or other RBAC management integration Deep knowledge of multi threading application or using framework like vertx Handle Project Banking/ Finance Full WFO at Sudirman & Gajah Mada

3.2. Job vacancy segmentation

The segmentation of job vacancies aims to group postings with similar content into coherent clusters, thereby facilitating the identification of distinct patterns of skill demand within the digital labor market. By organizing job postings into clusters, it becomes easier to highlight differences in required competencies across groups of occupations. This approach also helps reduce the complexity of analyzing large-scale text data, making the findings more interpretable.

To achieve this, the K-Means clustering algorithm was applied to the job posting data, which had been transformed into numerical feature vectors using the TF-IDF method. The similarity between documents was calculated based on the Euclidean distance of these TF-IDF vectors. The optimal number of clusters was determined using the elbow method, while the quality of the clustering was validated with the Silhouette Score and the Davies-Bouldin Index (DBI).

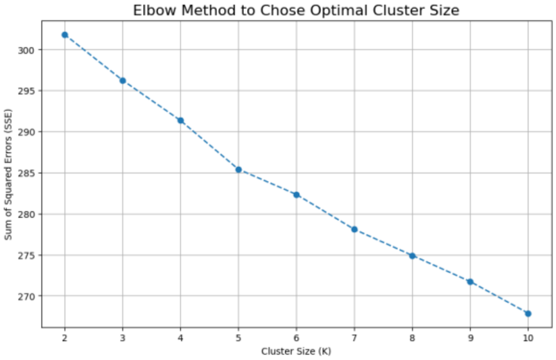


Figure 6. Elbow method chart.

Based on the elbow method graph (Figure 6), it can be observed that the decrease in Sum of Squared Errors (SSE) begins to slow significantly at K = 5, indicating the presence of an elbow point. This suggests that five clusters may serve as an efficient choice, as adding more clusters beyond this point yields only marginal improvements in reducing total error. However, to ensure the quality of the clustering results, two additional evaluation metrics—Silhouette Score and Davies-Bouldin Index (DBI)—were employed.

Table 3. Evaluation results.

k	Silhouette Score	Davies-Bouldin Index
2	0.021	6.500
3	0.026	5.498
4	0.025	5.211
5	0.032	4.656
6	0.030	4.759
7	0.035	4.592
8	0.037	4.178
9	0.038	4.289
10	0.043	4.052

Table 3 shows that the highest Silhouette Score and the lowest Davies-Bouldin Index are obtained at K = 10, indicating better separation and internal cohesion compared to other values. Although the SSE curve suggests that 5 clusters might be more efficient, dividing 350 job postings into 10 clusters (about 35 postings per cluster) is still reasonable and allows for a more detailed segmentation of competencies and job roles. Using only 5 clusters (around 70 postings per cluster) would risk oversimplifying the diversity of job types, whereas 10 clusters provide a more balanced and meaningful representation of the labor market structure.

Table 4. Cluster size and keywords for each cluster.

Cluster	Size	Keywords	Context
Cluster 1	42 documents	project, business, management, team, system, development, solution, user, requirement, project management	Project management and the development of technology-based business solutions.
Cluster 2	46 documents	database, oracle, system, program, sql, application, information, java, computer, engineering	Database management and application system programming.
Cluster 3	46 documents	security, network, infrastructure, server, cloud, linux, troubleshoot, support, monitor, technical	IT infrastructure and technical support for networks and servers.
Cluster 4	46 documents	skill, design, analytical, planning, excellent, problemsolving, communication, presentation, web, application	Application design as well as analytical and communication skills to support system development.
Cluster 5	27 documents	product, customer, sale, client, business, digital, management, scrum, data, user	Digital product development and customer service in business and technology contexts.
Cluster 6	66 documents	application, design, code, development, test, java, team, backend, spring, maintain	Backend application development, particularly using Java and supporting frameworks.
Cluster 7	26 documents	test, automation, test case, bug, quality, script, defect, mobile, tool, report	Software testing and test automation to ensure application quality.
Cluster 8	25 documents	data, model, pipeline, ETL, analysis, business, warehouse, machine, insight, python	Data science and data engineering.
Cluster 9	14 documents	audit, security, compliance, risk, control, policy, assessment, cybersecurity, ISO, governance	Information security auditing, regulatory compliance, and IT risk management.

Cluster 10	12 documents	website, search, engine, optimization, mobile apps, content, insurance, indonesia, financial service, hire	Digital marketing and mobile application development, including SEO and application-based financial services.

The clustering results reveal ten interrelated domains that reflect the multidimensional nature of digital and information technology work. Managerial clusters emphasize coordination and innovation, while technical clusters (backend development, data engineering, and IT infrastructure) form the core of digital operations. Integrative clusters focused on testing, cybersecurity, and analytics highlight the growing demand for quality assurance and risk control. The presence of digital marketing and compliance clusters further demonstrates the convergence of technology, governance, and business. Overall, these patterns indicate a labor market that increasingly values hybrid competencies, where technical expertise is complemented by managerial and analytical capabilities.

3.3. Competency profiles in digital labor market

Competency profiles in digital labor market extraction using NMF was conducted through two approaches, namely on the entire dataset and on each individual cluster. The first approach aims to identify a set of general competencies that frequently appear in job postings in aggregate, without taking into account segmentation by job type. Meanwhile, the second approach is carried out to uncover specific competencies that are prominent within each job group (cluster) formed through the clustering process using K-Means algorithm.

3.3.1. Topic modeling on the entire dataset

Before building the NMF model, an exploration of various topic numbers (K) was conducted to determine the most optimal number of topics. This process involved calculating the coherence score for topic numbers ranging from 2 to 10. Based on the evaluation results, it was found that the optimal number of topics is eight, with a coherence score of approximately 0.6435.

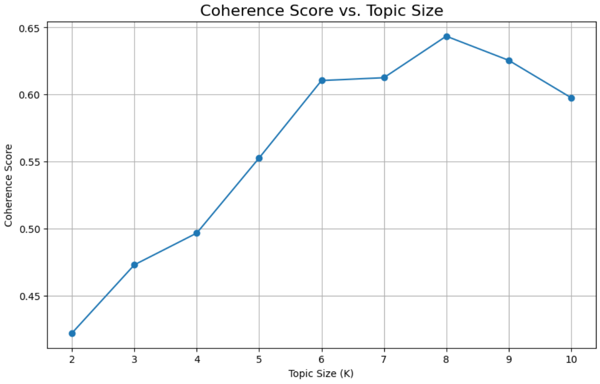


Figure 6. Coherence score for each number of topics

Based on the topic modeling results using NMF, eight main topics were identified, each reflecting a variety of competencies required in the field of information and technology. Each topic was derived from a set of dominant keywords extracted from job descriptions and minimum qualifications, and subsequently interpreted as a specific area of expertise. To facilitate understanding, a summary of the keywords along with a description of the corresponding competencies for each topic is presented in Table 5. These findings indicate that industry demands are not limited to technical skills such as software development, data engineering, and computer networks, but also encompass soft skills such as communication, project management, and information security—skills that are increasingly crucial in the digital era.

Table 5. Topic modeling for the entire dataset.

Topic	Keywords	Description
Topic 1	information, program, system, technology, computer, information technology, minimum, engineering, information system, science	Fundamental competencies in information technology, information systems, and computer science.
Topic 2	server, network, system, infrastructure, database, support, backup, cloud, management, troubleshoot	Expertise in network systems, servers, cloud, and IT infrastructure, including troubleshooting and technical maintenance.
Topic 3	product, project, management, business, team, client, technical, ability, stakeholder, customer	Managerial competencies such as project and product management, team collaboration, and communication with clients and stakeholders.
Topic 4	data, model, pipeline, etl, analysis, data model, warehouse, experience data, data analysis, business	Skills in data processing, data engineering, building data pipelines and data models for business analysis.
Topic 5	test, test case, automation, case, bug, quality, defect, script, automate, quality assurance	Competencies in quality assurance and software testing, both manual and automated, as well as software quality control.
Topic 6	skill, skill excellent, analytical planning, planning skill, planning, excellent problemsolving, skill good, communication presentation, presentation skill, presentation	Soft skills such as analytical thinking, planning, communication, and presentation, which are essential to support technical performance.
Topic 7	design, application, code, web, development, backend, user, frontend, developer, develop	Competencies in software and web application development, including both frontend and backend.
Topic 8	security, compliance, audit, risk, network, vulnerability, incident, assessment, policy, cybersecurity	Expertise in cybersecurity, auditing, risk management, and information security policies.

The results of the topic modeling identified eight main topics that reflect the skill requirements in job postings within the fields of information technology and data science. These topics include fundamental competencies in information technology and computer science, technical expertise in networks, servers, cloud, and IT infrastructure, managerial skills such as project and product management, as well as proficiency in data engineering and business analytics. Although topic modeling on the entire dataset successfully identified several core competencies commonly required in the information and technology sector, this approach has limitations in capturing the diversity present within specific job types. The generated topics remain largely aggregate in nature, tending to reflect general dominant trends that may obscure the distinct characteristics of more specialized job categories. Therefore, topic modeling was also conducted separately for each cluster produced by the K-Means grouping process, with the aim of uncovering more distinct and specialized competencies within each segment of job postings. This approach enables a more focused and contextual extraction of topics, as it is applied to data that has already been grouped based on similarities in content from job descriptions and qualification requirements.

3.3.2. Topic modeling by cluster

The results of topic modeling by cluster reveal that each cluster contains sub-topics that are consistent with specific job themes, covering both technical and non-technical aspects. For instance, Cluster 6, which is dominated by keywords such as “java,” “spring,” “backend,” “framework,” and “application”, clearly represents competencies in backend application development, particularly using Java and its supporting frameworks. This cluster corresponds to job roles such as Backend Engineer, Java Spring Developer, and Software Engineer, highlighting the strong demand for expertise in programming languages and system architecture. Additionally, sub-topics related to soft skills—such as communication and problem-solving—also emerge, indicating that interpersonal abilities remain a crucial component of the skill sets demanded in the information technology sector. This analysis reinforces earlier findings that the industry requires not only deep technical expertise but also cross-functional collaboration and strong communication capabilities. These examples illustrate how topic modeling, when applied to each cluster, provides more detailed and contextualized insights into the competencies required by different segments of the digital labor market. By analyzing clusters separately, it becomes possible to distinguish between technical specializations, such as backend development or data engineering, and broader professional competencies, such as project management or information security.

Table 6. Topic modeling for each cluster.

Cluster	Topic/Keywords	Relevant position
Cluster 1	Topic 1: business, user, test, requirement, solution, process, analyst, system, functional, specification Topic 2: project, management, project management, manage, risk, team, plan, work, schedule, client Topic 3: team, technical, development, design, system, software, technology, ability, environment, effectively	Business Analyst, Functional Analyst, System Analyst, Project Coordinator, Project Manager, Technical Project Lead, Solution Architect, Product Analyst
Cluster 2	Topic 1: program, system, information, test, software, work, application, development, net, minimum Topic 2: database, oracle, performance, backup, recovery, security, problem, design, oracle database, manage	Software Tester, QA Analyst, System Administrator, Oracle Developer, Database Engineer, IT Support Engineer, Application Support Analyst, Systems Integration Engineer
Cluster 3	Topic 1: system, skill, infrastructure, server, management, experience, manage, performance, user, team Topic 2: security, network, device, network security, security device, incident, design, vulnerability, experience, firewall	IT Infrastructure Engineer, Network Administrator, Cybersecurity Specialist, Systems Engineer, Technical Support Analyst, Security

Cluster	Topic/Keywords	Relevant position
	Topic 3: client, technical, maintenance, provide, troubleshoot, project, procedure, engineering, service, support	Operations Center (SOC) Analyst, Maintenance Engineer
Cluster 4	Topic 1: strong, good, communication, planning, skill excellent, process, skill good, problemsolving skill, excellent problemsolving, good communication Topic 2: design, web, user, developer, development, application, improve, cs, frontend, mobile	Digital Business Analyst, Product Owner, Scrum Master, Digital Strategist, Business Development Executive, Presales Consultant, Product Marketing Specialist, Stakeholder Engagement Lead
Cluster 5	Topic 1: digital, data, business, market, ability, design, drive, analysis, project, strategy Topic 2: scrum, product backlog, backlog, development, team, knowledge, collaboration, owner, product owner, stakeholder Topic 3: client, sale, customer, information, information technology, technical, solution, presentation, technology, system	Quality Assurance Engineer, Automation Tester, QA Analyst, Software Tester, SDET (Software Development Engineer in Test), Test Case Designer
Cluster 6	Topic 1: experience, system, technical, team, design, use, work, frontend, backend, understand Topic 2: io, swift, application, ability, performance, design, strong, test deploy, exist, experience Topic 3: java, spring, experience, development, spring boot, boot, application, framework, continuous, program Topic 4: skill, test, code, excellent communication, excellent, good, communication skill, application, program, unit test	Full-Stack Developer, Mobile iOS Developer, Java Spring Developer, Software Engineer, Backend Engineer, Frontend Developer, DevOps Intern, QA Automation Tester
Cluster 7	Topic 1: quality, defect, quality assurance, assurance, work, automation, query, test case, tool, strong Topic 2: application, security, web, web application, mobile, process, good, mobile application, program, computer Topic 3: use, system, script, project, tool, automate, performance, test script, user, integration	Quality Assurance Engineer, Automation Tester, Mobile Security Analyst, Application Security Tester, Test Engineer, Web App QA Analyst, Integration Specialist, Test Script Developer

Cluster	Topic/Keywords	Relevant position
Cluster 8	Topic 1: analysis, experience, business, analyze, data analysis, company, model, minimum, good, large Topic 2: pipeline, model, design, familiar, design develop, machine, develop, experience, docker, collaborate stakeholder Topic 3: etl, warehouse, data warehouse, requirement, ensure, skill, data management, experience, knowledge, management	Data Analyst, Data Engineer, Machine Learning Engineer, Business Intelligence Developer, ETL Specialist, Data Warehouse Consultant, Analytics Lead, AI Pipeline Developer
Cluster 9	Topic 1: audit, standard, internal, risk, control, system, work, internal external, external, risk assessment Topic 2: security, incident, security policy, policy, tool, vulnerability, risk, assessment, cybersecurity, requirement Topic 3: audit, cybersecurity, management, strong, plan, conduct, risk, control, knowledge, information system	IT Auditor, Cybersecurity Auditor, Risk and Compliance Analyst, Security Policy Officer, GRC (Governance, Risk, Compliance) Specialist, Internal Controls Analyst, Information Security Consultant
Cluster 10	Topic 1: mobile, work, apps, website, mobile apps, financial service, conglomerate, etc, insurance etc, work financial Topic 2: search, search engine, engine, optimization, engine optimization, team, strong, skill, minimum, research Topic 3: website, wordpress, understanding, strong, ability, skill, development, work, basic, project	Mobile App Developer, WordPress Developer, SEO Specialist, Digital Content Strategist, Financial Tech Support, Search Engine Analyst, Web Optimization Engineer, Insurance App Developer

To strengthen the validity of the topic modeling results in this study, a comparison was conducted with two relevant previous works, namely the study by Debao et al. [9] in China and Lukauskas et al. [3] in Lithuania. In Debao et al.'s research, TF-IDF vectorization and K-Means clustering were applied to job posting data specifically within the big data sector. Their clustering process resulted in ten job categories—such as big data engineer, data analyst, ETL engineer, and product manager—represented by keywords such as "development," "java," "database," and "analysis." This demonstrates consistency with the present study, particularly in the emergence of topics related to software development, data analysis, and product management. However, Debao et al.'s study was more focused on classifying positions based on job titles, while the current study emphasizes competence extraction based on the content of job descriptions and qualifications.

On the other hand, Lukauskas et al. utilized sentence-transformers (BERT) for text representation, followed by UMAP for dimensionality reduction and HDBSCAN for clusteringm [3]. Their approach yielded more dynamic job profiles that reflected specific skill demands, including both soft and technical skills identified through topic exploration and job advertisement segmentation. This study shares similarities in its use of semantic representation techniques and content-based segmentation. Nevertheless, unlike Lukauskas et al., who employed generative AI to construct job profiles, this research adopts a topic-based approach using Non-negative Matrix Factorization (NMF) and explores relevant job positions through dominant keyword interpretation.

4. Conclusion

This study implements a text mining approach to analyze competency segmentation in job vacancies within the information and technology sector using the K-Means algorithm for clustering and Non-Negative Matrix Factorization (NMF) for topic modeling. The data, obtained from the Kalibrr platform, consists of 350 job postings, which were preprocessed and numerically represented using the TF-IDF method. The clustering process resulted in 10 job clusters representing various fields such as backend development, data science, QA automation, digital marketing, and information security auditing. Furthermore, topic modeling identified eight key competencies needed in this sector, ranging from technical skills—such as programming, network systems, and cybersecurity—to non-technical abilities like project management and communication.

The analysis reveals that Indonesia’s labor demand is complex and multidimensional. Positions such as backend developers, data engineers, and cybersecurity specialists require both work experience and mastery of specific technologies. On the other hand, there is also significant demand for roles that combine soft skills and collaborative abilities, such as project managers and business analysts. These findings underscore the importance of education and training programs that not only focus on technical aspects but also emphasize interpersonal and managerial skill development. Moreover, segmentation based on job description content provides a sharper real-time view of the dynamics of digital labor market demand.

This research contributes to competency mapping in the digital workforce through a combination of clustering and topic modeling techniques. The findings can support educational and training institutions in designing curricula that better align with industry needs, particularly in areas such as software development, data analytics, and cybersecurity. For companies, the segmentation results can serve as the basis for job taxonomy design and competency-based recruitment strategies.

Future research should focus on enhancing semantic representation using deep learning methods such as BERT, Word2Vec, or FastText, which can better capture contextual relationships between technical and soft skills within job descriptions, allowing for more precise identification of emerging competencies. Spatial analysis can also be integrated to map geographic variations in skill demand across regions, enabling policymakers to design targeted workforce development initiatives and regional training programs. Additionally, incorporating variables such as salary range, education level, and work experience would enrich the analysis by linking required competencies to job value and qualification trends. The use of semi-supervised learning with external job label datasets is also recommended to strengthen validation and improve generalizability in future competency mapping studies.

References

[1] R. Shahnazi, “Do information and communications technology spillovers affect labor productivity?,” *Struct. Chang. Econ. Dyn.*, vol. 59, pp. 342–359, 2021, doi: https://doi.org/10.1016/j.strueco.2021.09.003.

[2] Z. Elmenzhi, S. El Fassi, R. Frij, I. A. Lhassan, and A. Maghni, “The impact of digital transformation on E-recruitment performance: An empirical study,” *Edelweiss Appl. Sci. Technol.*, vol. 9, no. 5, pp. 884–895, 2025, doi: 10.55214/25768484.v9i5.7039.

[3] M. Lukauskas, V. Šarkauskaitė, V. Pilinkienė, A. Stundžienė, A. Grybauskas, and J. Bruneckienė, “Enhancing Skills Demand Understanding through Job Ad Segmentation Using NLP and Clustering Techniques,” *Appl. Sci.*, vol. 13, no. 10, 2023, doi: 10.3390/app13106119.

[4] L. C. Moller, “Indonesia Economic Prospects : Boosting the Recovery : Indonesia Economic Prospects June 2021 (English),” Washington, D.C., 1385.

[5] Kompas, “Indonesia kekurangan 600 ribu talenta digital per tahun,” 2023.

[6] G. Graetz, P. Restrepo, and O. N. Skans, “Technology and the labor market,” *Labour Econ.*, vol. 76, p. 102177, 2022,

doi: https://doi.org/10.1016/j.labeco.2022.102177.

[7] V. B. Kobayashi, “Text analytics applications in job analysis and career research,” Amsterdam Business School Research Institute (ABS-RI), 2023.

[8] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proceedings of the 14th International Conference on Neural Information Processing Systems*, in NIPS’00. Cambridge, MA, USA: MIT Press, 2000, pp. 535–541.

[9] D. Debao, M. Yinxia, and Z. Min, “Analysis of big data job requirements based on K-means text clustering in China,” *PLoS One*, vol. 16, no. 8 August, pp. 1–14, 2021, doi: 10.1371/journal.pone.0255419.

[10] H. Djaya Siswaja and R. Tri Prasetio, “ANALISIS LOWONGAN PEKERJAAN DI BIDANG TEKNOLOGI INFORMASI BERDASARKAN LOKASI MENGGUNAKAN TEKNIK KLASTERING K – MEANS,” *J. Responsif Ris. Sains dan Inform.*, vol. 7, no. 1, pp. 19–25, Feb. 2025, doi: 10.51977/jti.v7i1.2020.

[11] E. M. Agustyani and I. Santoso, “Analisis Lowongan Pekerjaan Studi Kasus: Portal Lowongan Kerja Jobstreet,” *Semin. Nas. Off. Stat. 2019 Pengemb. Off. Stat. dalam mendukung Implementasi SDG’s*, pp. 1–10, 2020.

[12] V. Kumar and S. Priya, “Applying K-Means Clustering to Group Jobs Based on Location and Experience Level : Analysis of the Job Recommendation,” vol. 4, no. 3, pp. 178–189, 2024.

[13] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, and J. Teixeira, “Analysis of document pre-processing effects in text and opinion mining,” *Inf.*, vol. 9, no. 4, pp. 1–13, 2018, doi: 10.3390/info9040100.

[14] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[15] I. Yahav, O. Shehory, and D. Schwartz, “Comments Mining With TF-IDF: The Inherent Bias and Its Removal,” *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 437–450, 2019, doi: 10.1109/TKDE.2018.2840127.

[16] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.

[17] Z. L. Chen, “Research and Application of Clustering Algorithm for Text Big Data,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7042778.

[18] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm,” 2020, doi: 10.4108/eai.24-1-2018.2292388.

[19] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.

[20] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa,” *J. Sains dan Manaj.*, vol. 9, no. 1, p. 96, 2021, doi: 10.31294/evolusi.v9i1.10428.

[21] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Front. Sociol.*, vol. Volume 7-, 2022, doi: 10.3389/fsoc.2022.886498.

[22] V. Gallego, A. Freixes, and J. Lingan, “Applying Machine Learning in Marketing: An Analysis Using the NMF and K-Means Algorithms,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2025, pp. 14–26. doi: 10.1007/978-3-031-78238-1_2.

[23] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, and Q. Zheng, “Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling,” *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 2308–2314, Feb. 2017, doi: 10.1609/aaai.v31i1.10832.

[24] S. Mifrah, “Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID’19 Corpus,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5756–5761, 2020, doi: 10.30534/ijatcse/2020/231942020.

[25] S. Adhya, A. Lahiri, D. K. Sanyal, and P. P. Das, “Evaluating Negative Sampling Approaches for Neural Topic Models,” *IEEE Trans. Artif. Intell.*, vol. 5, no. 11, pp. 5630–5642, 2024, doi: 10.1109/TAI.2024.3432857.

[26] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, *Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks*, vol. 1, no. 1. Association for Computing Machinery, 2023.

[27] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proc. Natl. Acad. Sci.*, vol. 120, no. 30, p. e2305016120, 2023, doi: 10.1073/pnas.2305016120.