Big Data Analytics of Knowledge and Skill Sets for Web Development Using Latent Dirichlet Allocation and Clustering Analysis

Karina Djunaidi[1], Dine Tiara Kusuma[2]*, Rahma Farah Ningrum[3], Puji Catur Siswipraptini[4], Dina Fitria Murad[5]
Faculty of Energy Telematics, Institut Teknologi PLN, Jakarta, Indonesia[1, 2, 3, 4]
Information Systems Department-Binus Online Learning, Bina Nusantara University, Jakarta, Indonesia[5]

***Abstract*—Web development is a data-centric field and fundamental component of data science. The advent of big data analytics has significantly transformed the processes, knowledge domains, and competencies associated with Web development. Accordingly, educational programs must adjust to contemporary advancements by initially determining the abilities required for big data web developers to satisfy industry demands and adhere to current trends. This study aims to identify the knowledge areas and abilities essential for big data analytics and to create a taxonomy by correlating these competences with currently popular tools in web development. A mixed method consisting of semi-automatic and clustering methods is proposed for the semantic analysis of the text content of online job advertisements associated with the development of big data web applications. This methodology uses Latent Dirichlet Allocation (LDA), a probabilistic topic modeling tool, to uncover hidden semantic structures within a precisely specified textual corpus and average linkage hierarchical clustering as a clustering analysis technique for web developers. The results of this study are a web development competency map which is expected to help evaluate and improve the knowledge, qualifications and skills of IT professionals being hired. It helps to identify the roles and competencies of professionals in the company's personnel recruitment process; and meet industry skill requirements through web development education programs. The competency map consists of knowledge domains, skills and essential tools for web development such as basic knowledge, frameworks, design and user experience, database design, web development, cloud computing and other soft skills. Furthermore, the proposed model can be extended to several types of jobs in the IT sector.**

***Keywords*—Big data analytics; hierarchical clustering; Latent Dirichlet Allocation; web development; knowledge; skill**

I. INTRODUCTION

The revolution of Industry 4.0 carried out the concept of digitalization in all sectors producing big data. Big data consists of enormous volumes and a variety of data. It is impossible to manage and process the traditional management methods [1]. Big data analysis reveals hidden information, patterns, and correlations with new insights [2]. The valuable insights and implications derived from big data analytics are used in intelligent processes, such as guiding decision-making strategies in various organizations, including educational institutions, businesses, and governments. Big data are generated from many resources, such as websites, applications, emails, social media, and other multimedia platforms [3], [4], [5].

Websites as famous digital marketing in any organization, have caused increasing demand for data-oriented services, aligning knowledge and skill sets related to big data [2], [6]. Big data analytics is defined as the process of examining, processing, and analyzing large and complex data sets that cannot be handled by traditional methods. The goal of these analytics is to identify patterns, trends, and useful insights from structured and unstructured data [7] [8].

Recently, the technology lifecycle has shown a significant increase in big data-driven websites. Some modern products and services have been embedded in big data-oriented websites; thus, they have become an interesting topic for researchers[9], [10]. Digital transformation across business and industrial eras provides an exciting experience for software and service-based economics, for which modern websites can provide valuable information from big datasets [11], [12]. During this process, websites played a significant role in modernizing numerous sectors [13], [14], [15]. Web developers played an important role in developing ICT-based industries. Web developer is one of the information technologies (IT) occupations projected to grow by almost ten percent by 2033, it is much faster than other IT occupations [16]. The main task of web developers is to design and build a responsive website using popular programming languages such as HTML, cascading style sheet (CSS), and JavaScript. They are also responsible for testing, debugging, and integrating systems using application programming interface (API) services.

Big data causes new and challenging problems that need to be resolved using artificial intelligence. The Indonesian Ministry of Education, Culture, Research and Technology (Kemendikbudristek) stated that, only 15-20% of bachelor graduates have competencies match to their jobs. Skills and job profiles in the (Information Technology) IT sector is not clearly defined [17]. Therefore, some literature has discussed the big data of web developers and has become a hot topic among scientists in the last five years [18], [19]. Big data consists of approximately 5Vs; volume means a huge amount of data, variety means a type of data such as structured and unstructured data, velocity means high speed and real-time, veracity means reliable and accurate, and variability means volatility [1], [9]. These five characteristics of big data form the basis of the web development life cycle through methodologies and approaches in Big Data Analytics (BDA). This study aims to reveal hidden information and the value of implementing BDA in web

*Corresponding Author.

developers' occupations by identifying its knowledge domains and skill sets.

Given this context, BDA requires a diverse set of skills, programming languages, web development tools, and frameworks. The web development industry is a dynamic work environment that relies entirely on the resources of qualified people. The competence of BDA specialists strongly influences the quality of BDA-based products and services. BDA has grown in popularity, as has the requirement for qualification. A semi-automatic methodology was proposed to analyze collections of online job advertisements (ads). Our methodology is based on semantic analysis – hierarchical clustering (SA-HC) of BDA job ads using Latent Dirichlet Allocation (LDA) and average linkage hierarchical clustering. Latent Dirichlet Allocation (LDA). LDA is a generative statistical model used in a wide range of research in natural language processing and data analysis, such as topic modeling, sentiment analysis, and text analysis. This study revealed the core skills and knowledge required for BDA based on discovery topics, using LDA and hierarchical clustering analysis. The topics are mapped based on competency domains to reveal a structured taxonomy for BDA. Furthermore, the technologies required for BDA, such as programming languages, databases, and big data tools, are extracted.

The main contributions of this research are:

- A competency taxonomy for BDA developed by mapping the topics according to competency domain.
- The complex datasets provide a wide range of web developers topic area.
- A novel mixed method consists of latent Dirichlet allocation (LDA) and average linkage hierarchical clustering.
- BDA contributes significantly to decision-making processes because it has high granularity detailed information related to web developers' knowledge and skill sets.
- This research has been conducted by involving expert judgement in determining web development analysis.

II. Related Works

The methodology of this study provides a comprehensive explanation of big data analytics and semantics associated with them. It is also based on a content analysis of the textual content of BDA job ads using generative topic models to reveal the knowledge domains and skill sets required for BDA. Therefore, the background of the study is addressed under two subheadings: big data web developers and topic models / big data analytics.

A. Big Data Analytics

Big data analytics is the process of analyzing large volumes of documents to extract meaningful insights and values. High technology industries pioneered the method of deriving values from BDA. It includes a variety of data-intensive technologies that are capable of processing large volumes of data [1], [20], [21]. High-level management uses big data to impact grateful decision making, which is one of the parameters useful for BDA. This makes a substantial contribution to decision making because large-scale data contains specific information. The BDA consisted of five stages:

- Data retrieval refers to a set of text, images, videos from the Internet, sensors, and e-commerce; for example, social media generates billions of related data every day.
- Data acquisition consists of collecting data from sources, preprocessing to clean datasets, and transforming the data for purposes such as classification or clustering.
- Data management file system was created for effective data storage and processing of large datasets. The industry deploys big data cloud models and systems, such as the Hadoop distributed file system (HDFS) and NoSQL.
- Data analytics is the process of extracting insights from massive datasets using artificial intelligence techniques such as machine learning and data mining. BDA reveals knowledge for decision making by identifying hidden patterns, links, and interconnections. User experiences such as customer service and decision support can be enhanced by BDA.
- Data visualization is a graphical representation commonly used in big data. Researchers can use general software such as R studio or MATLAB to create some visualizations. Other than that, industries usually use their own applications, such as GIS-based 3D visualization, to monitor traffic data.

B. Web Developers

Web developers is one of information technology (IT) occupation which projected to grow almost ten percent up to 2033, it is much faster than other IT occupations[16]. The main task of web developers is to design and build a responsive website using popular programming languages, such as HTML, CSS, and JavaScript. They are also responsible for testing, debugging, and integrating systems using API services. Reliable and fast interconnection between web and mobile development has become a trending issue. Web developers can access WSs through application programming interfaces (APIs) in social media platforms [18]. Furthermore, web development tools have been implemented to enhance the accessibility of artificial intelligence for researchers and end-users [19].

C. Latent Dirichlet Allocation (LDA)

LDA, a generative statistical model, has become a popular method for topic modeling in text mining [2]. Latent pertains to the identification of semantic content in corpus documents through the analysis of the underlying semantic structures. The generative approach in LDA ensures the allocation of terms in a document to random variables, followed by semantic clustering through a repeating probabilistic process grounded in Dirichlet distribution. LDA is an unsupervised learning methodology that does not require labeling or training datasets. It can be concluded that LDA can be efficiently applied to large corpus documents to identify semantic patterns. In the past five years, LDA has gained popularity in text mining studies spanning a variety of contexts, including e-commerce reviews, natural language processing, information extraction, sentiment analysis, and

social media trend analytics. Likewise, this approach has been used as a successful strategy in some studies that analyzed online job advertisements from businesses and industries. In the past, topic models were only created for textual data analysis but are currently being applied to a variety of data sources, including genetic data, photos, videos, and social networks. For these reasons, this study implemented LDA as a topic modeling method.

D. Average Linkage Clustering Analysis

Unsupervised learning such as hierarchical clustering, has become a popular method in data analysis. A superior cluster quality was provided by hierarchical clustering, thereby diminishing the sensitivity of clustering to various problem types. It comprises two methodologies: bottom-up and top-down. Agglomerative as a bottom-up approach initially treats each instance as an individual cluster, which is subsequently merged to form bigger clusters; this is known as Average-Linkage Hierarchical Clustering (ALHC) [22]. This process continues until all the clusters are combined into a single giant cluster containing all the instances. The hierarchical clustering method identifies common traits and job profiles in IT job posts, including competency, programming languages, web development tools, and frameworks [23].

III. Research Method

This study analyzed the content of web developer's job advertisements. The proposed research methodology is described in Fig. 1 which consists of three main phases: data collection, text preprocessing, and LDA implementation. The following figure illustrates the overall process.

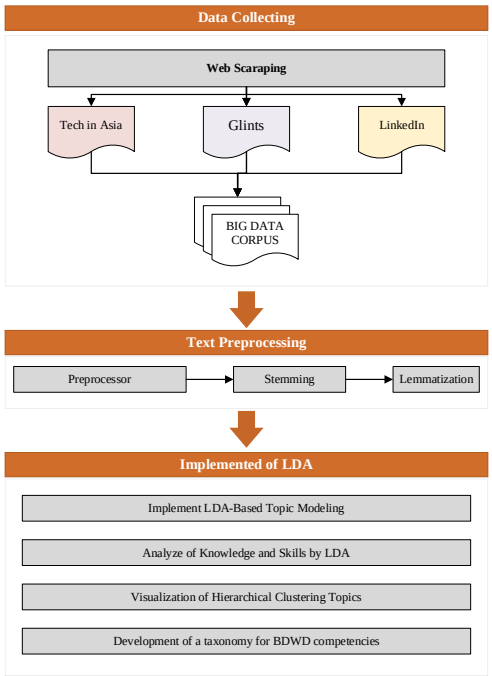

Fig. 1. Research methodology of big data analytics using LDA and clustering analysis.

A. Data Collection

The data used in this study were obtained from online job advertisements published by Glints [24], Tech in Asia [25], and LinkedIn [26]. A total of 2649 data were collected using the web scraping technique from January 2023 to July 2024. Job advertisements were searched using the keyword 'web developer', 'web development', and some expertise fields, including job title, job description, and required skills were collected. Table I presents the sample dataset.

TABLE I. Sample of Data Collection

Job Title	Job Description	Required Skills
Full stack Developer	Design, build, and optimize front-end and back-end code. Develop and maintain web applications. Integrate third-party APIs and services. Perform testing and debugging to ensure application quality. Collaborate with the design team to implement a responsive and intuitive user interface. Manage databases and perform query optimization. Provide continuous technical support and bug fixes. Keep up with the latest technological developments and implement best practices in development.	jQuery Debugging Laravel JavaScript CSS3 GIT SQL PHP HTML5
Web Developer	Develop new web application or customize existing application Learn new technology when required in the process of application development Problem solving and working with team on a project	Node.js REST API Laravel PHP

B. Text Preprocessing

Text preprocessing has become a crucial stage in information retrieval research. However, text data often comes in an unstructured form and is full of noise, especially when obtained from sources such as social media or websites. Therefore, text preprocessing is a crucial initial step in aligning data before being directed to further stages of analysis. The preprocessing stage plays a major role in removing text data from noise, which can damage the quality of the results [2], [27], [28], [29], [30]. Text preprocessing applied to the experimental data set consisted of several sequential stages. First, the text data were divided into words (tokens/parsing), known as tokenization, to obtain meaningful attributes [31], [32].

Tokenization divides text in the form of sentences or paragraphs into tokens/parts that are then represented by data vectors [33], [34], [35], [36]. Furthermore, web links, personal tags, and characters/affixes with no meaning were removed. The next stage was the stop word process. Stop words are used to reduce the number of words in a document, which affects the speed and performance of Natural Language Processing (NLP) [37], [38], [39].

The text preprocessing stages involved in converting textual data into keywords in WordStat ver.2024 are:

1) Pre-processor: This option allows custom text transformations to be analyzed before or instead of the execution of the other three standard processes: lemmatization, exclusion, and categorization. These transformations are achieved by executing specially designed external routines that

can be accessed in the form of Python scripts, external EXE files, or functions in dynamic link library.

2) Stemming: Stemming is a process used in text preprocessing to convert words into their base form [40], [41], [42]. For example, the word 'running' is changed to 'run. This helps in text analysis because it reduces the variation in words with similar meanings. Stemming ignores suffixes and prefixes to arrive at the base form of the word. Its use is common in applications such as information retrieval, sentiment analysis, and text mining. This can be useful for improving the accuracy of text analysis.

3) Lemmatization: Lemmatization is a process in text preprocessing that aims to change words into their basic form or 'lemma'. Unlike stemming, which only cuts off the endings of words, lemmatization considers context and changes words into their grammatically correct basic forms [43], [44]. After the preprocessing stage was completed, each text (job ad) in the dataset was defined as a word matrix. As a result of the preprocessing, the word space size for the entire dataset was reduced from 28232 to 22773. The dataset consisting of the job ads "TGL Web Developer and Digital Designer" is characterized by 22773 unique words, which also refers to the word matrix size for each ad. The number of matrices/vectors is 1868, which is also the number of job ads. The Document Term Matrix (DTM) created for this analysis consists of 1868 rows and 22773 columns. In other words, the DTM shows that 1868 job ads were represented by a word space consisting of 22773 terms. The DTM weighting process is performed by considering the word frequency.

C. Implementation of LDA-Based Topic Modeling

This step of the experimental analysis entails the application of a topic model to the dataset to reveal the domain knowledge and expertise necessary for BDA in a clear and comprehensible manner. The LDA model is a document generation model based on Bayesian theory, that excels in extracting themes and features from voluminous texts. This concept has found extensive application in fields such as text mining and information retrieval [45]. The inherent qualities of research that utilizes semantic analysis of job advertisements contribute to the effectiveness of LDA as a topic model [2]. LDA-based topic modeling assumes that the distribution of subjects in texts and the distribution of words within topics are mutually independent. Identical terms may manifest at varying degrees across distinct subjects. Likewise, specific subject matter may manifest to varying degrees in several written materials. The fundamental premise of the LDA model is derived from the Bayesian joint probabilistic model. The objective of this study is to use LDA-based topic modeling to reveal the underlying semantic structures (word clusters) in a textual corpus of job advertising.

Once the LDA model was implemented, the probability distribution for each topic is computed using Bayesian estimation methods in conjunction with the Dirichlet distribution. The WordStat ver.2024 tool was utilized to implement an LDA model in this experimental investigation [46]. This tool is specifically designed to implement an LDA model. WordStat was implemented with varying iteration counts and was stabilized after 500 successive Gibbs sampling iterations.

The Bayesian inference model is the most important component of the LDA model [45], which is produced by the three-layer Bayesian probability of the "topic word" in the text. The diagram in Fig. 2 illustrates the topological structure of LDA. The fundamental algorithms of the data of a data-sampling algorithm and a feature-weight algorithm. The data samples in this study were collected using a web scraping technique.

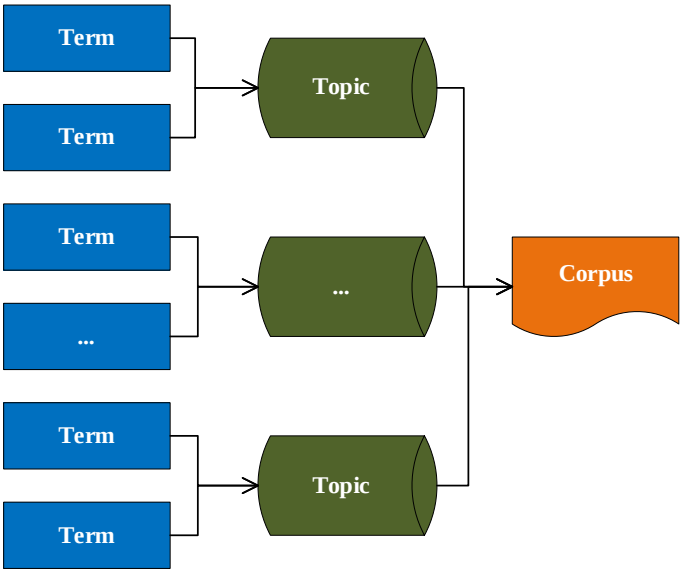

Fig. 2. Topological structure of LDA.

Within the LDA model, researchers explicitly allocated topic names to the identified themes, in accordance with the descriptive keywords. The naming of topics is based on the significance of all keywords and involves professional experts in the field of web developers through Forum Group Discussions (FGD). Therefore, the topic titles used may differ depending on the perspective of each researcher. LDA is an unsupervised generative probabilistic technique that is used to model a corpus.

Fig. 3 shows a diagram illustrating the LDA algorithm. The parameters used were as follows:

α and β are parameters of the previous distribution of θ. z is the designated theme for the nth word in the document count

ϑ is distribution of terms in number of topics theme

w is word in document

For the specified parameter θ, the mathematical equation to compute the probability distribution of the topic in Eq. (1) is as follows:

$$p(z|\theta) = \prod_n^N p(z|\theta) = \prod_{k=1}^K \theta_m^n \quad (1)$$

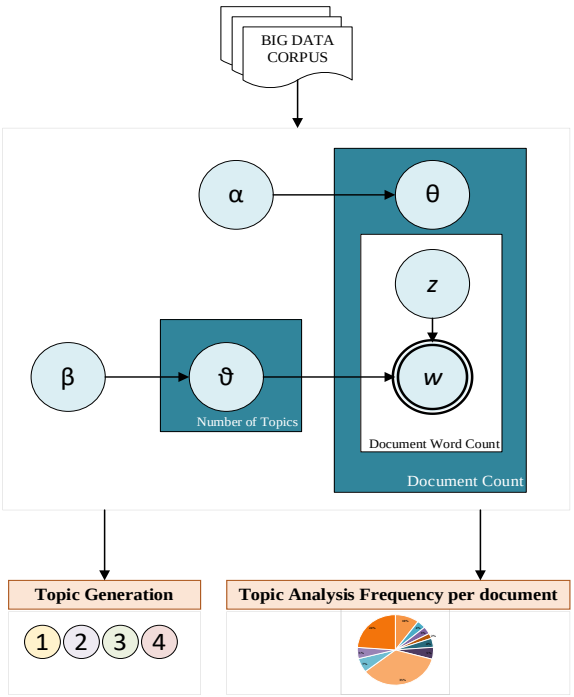

Fig. 3. Diagram illustrating the LDA algorithm.

Assuming the term-document matrix is defined and considering the number of documents comprising a corpus, these matrices often exhibit a significant size. Hence, it is customary in text mining to exclude sparse terms, which have a very low inclusion rate in documents. Typically, such an approach allows for a substantial reduction in the size of the matrix while preserving its essential relationships.

IV. RESULTS

BDA is presented to identify the core competencies of big data web development. First, the corpus document classifies the topics into skill sets. Then, the competency domain is mapped using these skill sets. Finally, most in-demand tools for BDA, programming languages, web development tools, and frameworks were analyzed to identify higher quality competencies. The results of the analysis are presented and discussed below.

A. Analyze Knowledge and Skills Using Latent Dirichlet Allocation

The corpus document formed by top three job advertisements comprised a wide spectrum of knowledge, skills, and job descriptions in the web development area. These spectra extended the coverage of the discovered topics of BDA. Three variables of job advertisements were used to combine the LDA-based topic modeling. Knowledge domains and skill sets of BDA were revealed and discovered from 26 trending topics with optimal granularity. As presented in Table II, topics were combined using descriptive LDA keywords and topic rates. The descending order and percentages are listed in Table II. Descending order means that the first term was the most occurrence and the last term was the least occurrence in a topic. The names of the discovered topics were automatically assigned using WordStat ver. 2024.

TABLE II. DISCOVERED TOPICS

TOPIC NAME	LATENT DIRICHTLET ALLOCATION KEYWORDS	RATE %
Cascading style	cascading style sheets; responsive web design; css; front end; web development; client requirements;	10.04
Programming language	programming language; python; typescript; javascript; kotlin; golang	9.32
Graphic design	graphic design; corel draw; adobe photoshop; pattern; teamwork; management; communication; marketing social media	8.34
HTML CSS PHP	php; html; css; doctrine; laravel; mysql; jquery; bootstrap; wordpress; framework	7.04
Graphic design video	video; editing; graphic; design; illustrations; canva; adobe; video editing; photo editing; image editing; motion graphics; multimedia design	5.44
Digital marketing	digital marketing; marketing strategy; online marketing; product marketing; sales and marketing; creative writing; creative design; google ads; content marketing;	4.90
Code review	code review; code integration; development; system development	4.85
Javascript	javascript; react js; laravel node; angular js; vue js; tailwind css; node js; cassandra; development javascript;	4.76
Big data	big data; data engineering; scala; apache spark; data mining; olap cubes; data cubes	4.63
Design tools	After effects; adobe photoshop; adobe illustrator; canva design; google sketch up; auto cad; google sketch up; interior design;	4.35
Relocation provided	relocation provided; lead data engineer hadoop apps; staff data engineer; engineer data platform;	4.30
Performance tuning	performance tuning; testing; development; skills; continuous delivery; clean coding	4.06
Full stack	full stack developer; full stack engineer; full stack; full stack web developer; front end; back end;	3.55
Database design	database design; data architecture; sql; sql server; modeling; postgresql; nosql; stored procedures; database development; query optimization;	3.35
Business requirements	business requirements; problem solving; metrics driven; operational excellence; application testing; agile development; coding standards;	3.19
Object oriented	object; oriented; oop; object oriented; object oriented programming	2.66
Interpersonal skills	skills; analytical; interpersonal; solving; problem; communication; administration; systems interpersonal skills;	2.51
User experience	user experience; user research; architecture; large language models; requirements gathering; user interface design; debugging	2.32
Search engine	search engine; search engine optimization; digital marketing; google analytics; google ads; instagram	1.57
API	rest api; rest api laravel; git	1.50

TOPIC NAME	LATENT DIRICHTLET ALLOCATION KEYWORDS	RATE %
Full stack developer	fullstack; programmer; developer; senior; engineer; backend; web programmer; angular developer; senior full stack developer;	1.46
Model view controller	model view controller; mvc; asp.net; sql; programming	1.44
Web service	amazon; service; web; aws; amazon web;	1.37
System UI	system; ui; administration; linux; ux; application; test; testing; mobile; integration	1.23
Online advertising	online advertising; paid advertising; google ads; instagram ads; digital marketing; market analysis	1.09
Information technology	information technology; service; communications; customer; security; management; product; service level agreements;	0.73

Table II shows that cascading style, programming languages, and graphic design were among the competencies with the highest demand in the BDA industry. Other knowledge and skills in the top ten were html css php, graphic design video, digital marketing, code review, JavaScript, big data, and design tools. The discovered topics also covered various emerging trends, such as database design, business requirements, object oriented, user experience, search engines and soft skill areas such as interpersonal skills which shed light on the priorities and demands in the ever-growing BDA industry.

B. Knowledge and Skill Mapping According to Competency Domains

This stage focuses on categorization and presents knowledge and skills in a structured manner. First, a mapping process was performed by associating knowledge and skills with the competency domains and workflows. Second, the knowledge and skills revealed by 26 topics were mapped into ten core competency maps developed for BDA.

Table III presents the distribution of knowledge and skills according to the competency map and their respective percentages. As presented in Table III, the first three competency areas are related to the function of web development, which consists of big data products, roles, and specialized web developers. The total rate of these competencies as the most important focus in web development area was 48.49%. The next five competency areas were related to the major discipline, comprising databases, web development frameworks, tasks, programming languages, and web development tools. The total rate for these competencies areas was 38.81%. The last two concern the interdisciplinary areas, consisting of educational background and soft skills, at a rate of 12.7%. These ten competency areas are discussed in detail below.

The first competency area, big data products in web developer area (8.96%). It contains five knowledge and skill items: digital marketing, social media, search engine optimization, data engineering, and data science. The second, role (33.92%) means a web developer must perform, such as web development, graphic design, software development, data engineering, etc. The third, specialized web developer, means specialization on web developer title (5.61%), contains some items of front-end developers, full stack developers, Java developers, and others. The fourth, databases (2.1%) as a query language and data storage area, has the top five highest demands in BDA: SQL server, MySQL, MongoDB, Oracle, and powerBI. Fifth, web development frameworks (2.42%) used to create interactive and progressive user interfaces, consisted of six items: Apache spark, Laravel, angular js, etc. The sixth, task (10.52%), has the duties of web developer comprising user interface design, application testing, user experience, video editing, technical, content creation, and object-oriented. The seventh, programming languages (6.4%), has a lot of items, HTML CSS JavaScript, Python, typescript, scala, and VBScript. The eighth, web development tool (17.37%), contained a variety of frameworks, programming languages, and software. The ninth, educational background (4.41%), comprised four majors: software engineering, computer science, information technology, and electrical engineering. Finally, soft skills (8.29%) included problem solving, public speaking, creative design, positive attitude, communication skills, time management, analytical skills, and critical thinking.

TABLE III. COMPETENCY MAP

ID	Competency Areas	Knowledge and skills	Rate %	Total %
1	Big data product/output	Digital marketing	5.69	8.96
		Social media	2.19	
		Search engine optimization	0.56	
		Data engineering	0.29	
		Data science	0.23	
2	Role	Web development	6.79	33.92
		Graphic designer	6.67	
		Software development	5.56	
		Back end	4.31	
		Full stack	2.87	
		Front end	2.77	
		End developer	2.29	
		Application development	1.52	
		Data engineer	1.13	
3	Specialized web developer	Web developer	2.45	5.61
		Frontend developer	0.96	
		Full stack developer	1.13	
		Java developer	0.39	
		Net developer	0.38	
		Software engineer back end	0.31	
4	Databases	SQL Server	1.33	2.1
		MySql	0.58	
		MongoDb	0.08	
		Oracle Database	0.07	
		Power BI	0.04	
5	Web development framework	Model view controller	0.37	2.42
		Apache spark	0.19	
		Php-laravel	0.89	
		Javascript frameworks	0.37	
		Angular js	0.15	
		Java spring	0.45	
6	Task	User interface design	2.68	10.52
		Application testing	1.25	
		Video editing	1.18	
		User experience	1.12	
		Technical	1.07	
		Cascading style sheets	0.94	
		Data analytics	0.86	
		Object oriented programming	0.72	
		Content creation	0.68	
7	Programming Languages	Html css javascript	3.98	6.4
		Asp net	0.99	

ID	Competency Areas	Knowledge and skills	Rate %	Total %
		Typescript	0.36	
		Development java	0.32	
		Scala	0.18	
		Visual basic	0.18	
		Php rest api	0.16	
		Python	0.08	
		Query languages	0.07	
		Vbscript	0.05	
		Server side	0.03	
8	Web Development Tools	Adobe photoshop	9.83	17.37
		Programming language	1.48	
		React js	1.36	
		Rest api	0.88	
		Amazon web	0.85	
		Node js	0.84	
		Spring framework	0.80	
		Corel draw	0.70	
		Web applications	0.64	
9	Educational background	Software engineer	2.05	4.41
		Computer science	2.08	
		Information technology	0.23	
		Electrical engineering	0.04	
10	Soft skills	Problem solving	1.69	8.29
		Public speaking	0.73	
		Active listening	0.72	
		Creative design	1.78	
		Positive attitude	1.14	
		Communication skills	1.35	
		Time management	0.35	
		Analytical skills	0.29	
		Critical thinking	0.23	

C. Identification of the High Demand Tools for BDA

Recently, collective environments of web development, a wide range of tools and technologies, such as programming languages, web development tools, and frameworks are used simulant. The corpus document was analyzed using a keyword indexing technique to reveal the tools and technologies required for BDA [29]. The findings of this analysis were divided into three main categories: programming languages, web development tools, and frameworks, which are discussed in detail in the following sections.

1) Programming languages: Programming languages are essential tools for application development and serve various applications. The job advertisement dataset was analyzed using keyword indexing to identify programming languages used in BDA. Table IV shows the top 12 programming languages required for BDA along with their percentages.

TABLE IV. PROGRAMMING LANGUAGES

Programming languages	Rate %
HTML CSS Javascript	59,2
Asp.Net	14,7
Typescript	5,31
Development Java	4,7
Scala	2,73
Visual Basic	2,73
PHP	2,43
Golang	1,56
Python	1,21
Query Languages	1,06
VBscript	0,7
Server Side	0,46

According to the results in Table IV, HTML CSS JavaScript is the superior programming language in this field, followed by Asp. Net and Typescript. The total percentage of these three programming languages was 73.9%, a high percentage that shows their superiority. HTML CSS JavaScript shows the most superior and widely used programming as evidenced by the percentage produced is 59.18%, more than half of the existing value. In addition, the Server-Side programming language currently appears to have the least used trend in data science in recent years.

2) Web development tools: Table V shows Web development tools are often used in conjunction with programming languages to develop software applications more easily. These tools contain various types of utilities, such as frames, libraries, and applications. As seen in Table V, Adobe Photoshop, as a tool that can be used for UI / UX development in Web Development. React JS is a tool that has a JavaScript library used to build user interfaces, is in second place, followed by Rest API, a tool used to build web services so that applications can communicate with each other using the HTTP protocol. The fourth is Amazon Web, a cloud computing platform that provides various services for the development, hosting, and management of web applications and digital infrastructure. Almost the same rate value Node js which is a JavaScript-based runtime environment allows developers to run JavaScript code outside the browser, usually on a server. WordPress is a web development tool that is rarely used, as can be seen from the small percentage of tool use in Table V.

TABLE V. WEB DEVELOPMENT TOOLS

Web Development Tools	Rate %
Adobe Photoshop	48,39
React Js	6,68
Rest API	4,32
Amazon Web	4,17
Node Js	4,12
Spring Framework	3,92
Corel Draw	3,46
Web Applications	3,16
Search Engine	3,06
Graphic Illustrators	2,71
Canva Design	2,11

Web Development Tools	Rate %
Vue Js	2,11
Google Sketch Up	1,91
Net Framework	1,61
Mobile Application	1,41
Auto Cad	1,15
Data Cubes	1
Microsoft Azure	1
Java Virtual Machine	0,85
Query	0,8
Xml	0,75
Development Git	0,4
Cloud Computing	0,35
Apache Kafka	0,3
WordPress	0,25

3) Framework: According to Table VI, the most popular frameworks are PHP Laravel, Java Spring, Model View Controller, and Java Script, with 66.33%. Apache Spark, Angular Js, React Js, Next Js, Vue Js, and Express Js are in the middle position of their usage, with a total of 23.2%. Ruby on Rails has been the least utilized framework in recent years.

TABLE VI. FRAMEWORK

Frameworks	Rate %
Php Laravel	28,43
Java Spring	14,38
Model View Controller	11,76
Javascript Frameworks	11,76
Apache Spark	6,21
Angular Js	4,9
React Js	4,25
Next Js	3,92
Vue Js	3,92
Express Js	3,92
Redux Js	1,96
Angular Angularjs	1,96
Php Jquery	1,63
Ruby On Rails	0,98

4) Visualization of hierarchical clustering topics: Fig. 4 shows the relationship between various skills and topics related to web developers. The dendrogram shown in Fig. 4 is a graphical representation of hierarchical clustering. The dendrogram in Fig. 4 helps to visualize the relationship between skills or abilities related to the topic of the web developer in a hierarchical manner. For example, Cascading Style Sheets (CSS), HTML, and JavaScript are grouped together because they are closely related to the development of web interfaces. Technically, JavaScript, CSS, HTML and Frameworks such as Angular and Spring are grouped more closely because these skills are often used together in web application development. In addition, the analytical skills and skills in the figure above have long branches, indicating greater differences from the other groups. In the Soft Skills grouping, Communication and Interpersonal skills were in a different group from technical skills, indicating that soft skills are important and conceptually different from technical web development skills.

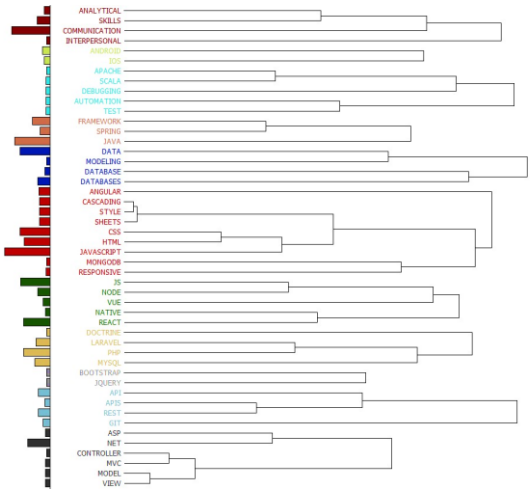

Fig. 4. Dendrogram agglomeration order of web developer and digital designers.

The visualization can be presented in a word cloud format in addition to being in a hierarchical diagram. Word Cloud is a method for visualizing text [47], [48], [49], [50]. This technique was used to identify and highlight the most frequently occurring words, thus providing insight into the dominant themes or topics in the text [51].

Fig. 5. Word cloud for web development.

Fig. 5 shows that the term 'Digital Marketing' is often paired with 'Web Development' as they complement each other in building and promoting an effective online presence. Some words that are often paired with 'web development' include various technical and non-technical aspects of web development. Here are some examples: HTML, CSS, JavaScript, Frameworks: Such as React, Angular, and Vue for the front-end, and Node.js and Django for the back end. Responsive design and Search Engine Optimization (SEO) are often paired with web development. Web Development is also often called Web Programming or Website Development or Web Application Development or Front-end Development or Back-end Development or Graphic Designer or can also be called Full-stack Development.

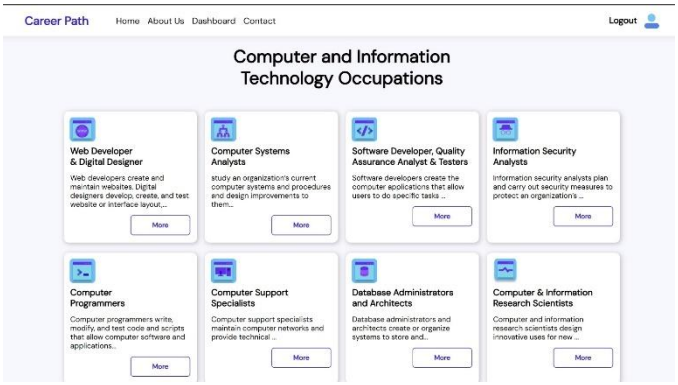

Fig. 6. Online dashboards of computer and information technology occupations.

As shown in Fig. 6, online dashboards of computer and information technology occupations has been deployed based on the ReactJS framework, and the Golang programming language.

V. DISCUSSION

This study has revealed the competencies for big data analytics over web development. First, the skill sets arranged by topic were identified from the data sets created using corpus documents of online job advertisement. Then, these skill sets were mapped into competency domains. Based on these results, the following ten competencies were identified:

- Big data product/output
- Role
- Specialized web developers
- Database
- Web development framework
- Task
- Programming languages
- Web development tools
- Educational programs
- Soft skills

The results show, in the big data web development field, HTML CSS Javascript, Asp.Net, and Typescript are the most demanded programming languages; Adobe Photoshop, React Js, Rest API presented as the most demanded programming tools; PHP Laravel, Java Spring, Model View Controller are listed as the most demanded databases. The results of this study have important implications for web developers' programs, which are summarized below.

A. From Big Data to Data Science

Big data is related to data science which has characteristics such as volume, velocity, veracity, variability, and variety. Competence in big data requires a wide range of knowledge and skills. Such as in Table II, volume and variety of data are crucial factors in big data product/output with a contribution of 8.96%. This includes skills in digital marketing, social media, and search engine optimization and data engineering, which reflect the need for expertise in managing and analysing big data for various purposes, such as digital marketing and search engine optimization. The competence of each role of each person involved in web development work needs to adjust data with high speed and variety to support the development of more innovative products, this is related to the characteristics of big data velocity and variety with a role competency area of 33.92%. In addition, the ability of data to be relied on accurately is related to the database which includes skills in using SQL Server, MySQL, Mango DB and Oracle which contribute 2.1%. This ability is especially important in maintaining veracity which is the main basis for precise analysis and accurate data-based decisions. Big Data Product/Output and Web Development Framework, with a total contribution of 11.38%, indicate that data generated from various sources and having various types can be optimized for various analysis and processing needs. This reflects the existence of variability in data, which allows flexibility in managing and applying data for different analytical purposes. As illustrated in Fig. 5, hierarchical clustering is one of the data science analysis methods that groups similar abilities to illustrate the relationship between big data and data science.

B. The Extensive Range of Knowledge Areas and Competencies

The Web Development industry is one of the most in-demand and fastest growing professional fields worldwide. It has a highly dynamic and competitive work environment with an ever-increasing, changing, and evolving demand for knowledge, skills, and abilities. The global workforce will be impacted by the adoption of AI, automation, and Big Data Analytics (BDA) [21]. Our analysis reveals the knowledge and skill domains that are in high demand for Big Data Analytics (BDA). The analysis findings indicate that expertise in Big Data Analytics (BDA) requires a broad spectrum of highly varied and interrelated knowledge, skills, and ability domains. It involves collecting, storing, processing, and analysing large amounts of data to generate insights that can be used for decision making. Taking these findings into account, a conceptual competency map is proposed to organize these knowledge and skills. The map consists of the following ten competency domains: big data products/outputs, roles, specialized web developers, databases, web development frameworks, tasks, programming languages, web development tools, educational background, and soft skills (see Table III).

The competencies found indicate that BDA expertise has an interdisciplinary background that requires the integration of a broad set of technical and non-technical skills. Although competency priorities vary from position to position, employers in the BDA industry generally demand a set of technical and non-technical skills, defined as the job skill set. In this regard, our analysis results offer a more comprehensive perspective for BDA employers to identify the job skill set required for effective candidate assessment. The knowledge domains and skill set also indicate the need for a demand-driven educational background approach based on interdisciplinary collaboration to achieve a competency-based web development curriculum. Our findings are also in line with industry reports and academic research that emphasize the use of technical and non-technical skills together based on an interdisciplinary background containing data science, web development, software engineering, computer science, mathematics, business science, statistics, and communication science.

C. Reconciling Hard Skills and Soft Skills for Web Developers

Soft skills and hard skills have important roles in different types of jobs, although their roles can differ depending on the field and position. Hard skills are technical and specific skills that can be measured and are usually acquired through education or training. Technical skills in web development such as mastery of programming languages, tools, or frameworks. Professional qualifications such as certifications or licenses required for a specific job. Soft skills are interpersonal and character skills that are harder to measure but are essential for success in the workplace. Soft skills relating to interpersonal capabilities such as problem solving, analytical thinking, and communication. Soft skills depend on regular activities and organizational experiences of people [52], [53]. Problem solving means the ability to think critically and find solutions to problems that arise. Communication is the ability to convey ideas clearly and listen to others. This study's findings, general soft skills required for BDA specialists are highly recommended to whom it may concern. The findings related to soft skills include creative design, problem solving, and communication as the most favorite soft skills needed by industries (see Table III). This perspective has total rate of soft skills is approximately 8,29% in all topics. In many jobs, soft skills such as empathy and communication facilitate good cooperation in a team, while hard skills ensure that technical tasks are completed.

D. Insights into the Use of Tools and Technologies

Web development involves a variety of tools and technologies to create, manage, and maintain a website. This study proves several aspects of its use, including frameworks, databases, APIs, and responsive design. Web developers use tools and technologies consisting of programming languages, tools and frameworks in developing web applications. The selection of these tools is tailored to the needs and latest developments in the web development industry. The results of corpus data on BDA show that HTML CSS, ASP.Net, and Typescript are the most widely used programming languages in this field. Adobe Photoshop, React Js, and rest API are web development tools that are in high demand [54], [55]. Finally, this study proves that the most widely used web development frameworks are PHP Laravel, Java Spring and Model View Control.

VI. Conclusion, Limitation, and Future Works

The finding of this study is big data analytics can clearly identify the industry needs of web development areas. A brief taxonomy of web development includes programming languages, databases, and web development tools that can improve the knowledge of students or job seekers in this field. LDA and hierarchical clustering algorithms prove that web developers can improve the innovation of businesses or organizations through digital marketing strategies. This study captures the updated industrial needs of the knowledge and skills of web developers because the datasets have been collected and managed in a wide and rigorous manner. The taxonomy of BDA competencies and skill sets was justified by three web developers' professionals through forum group discussions.

This study has several limitations. The first limitation comes from data collection using the web scraping technique; the text of job advertisement is biased because it is not specific enough or does not list relevant skills that are needed for the web developer's area. Second, the software used in the text preprocessing stage did not clean the data properly. There are some terms (stop words) still appear such as 'and', 'in', and 'to', so we must delete it manually. Third, the recurrence of phrases as synonyms forces us to determine the threshold for the most prevalent terms in job advertisements as 60.

In future work, some potential areas can be improved, such as comparing open-source tools, and Python to mine data. The latent semantic analysis (LSA) approach can be implemented to calculate the accuracy as a validated model.

Acknowledgment

This research was supported by the Ministry of Education, Cultural, Research, and Technology of Republic Indonesia and Institut Teknologi PLN based on Agreement Grant No. 0459/E5/PG.02.00/2024.

References

[1] P. V. Thayyib et al., "State-of-the-Art of Artificial Intelligence and Big Data Analytics Reviews in Five Different Domains: A Bibliometric Summary," Sustainability (Switzerland), vol. 15, no. 5, 2023, doi: 10.3390/su15054026.

[2] F. Gurcan and N. E. Cagiltay, "Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling," IEEE Access, vol. 7, pp. 82541–82552, 2019, doi: 10.1109/ACCESS.2019.2924075.

[3] B. Kumar, S. Roy, A. Sinha, C. Iwendi, and Ľ. Strážovská, "E-Commerce Website Usability Analysis Using the Association Rule Mining and Machine Learning Algorithm," Mathematics, vol. 11, no. 1, 2023, doi: 10.3390/math11010025.

[4] P. E. Justin Zuopeng Zhang, Praveen Ranjan Srivastava, Dheeraj Sharma, "Big data analytics and machine learning: A retrospective overview and bibliometric analysis," Expert Syst Appl, vol. 184, 2023, doi: https://doi.org/10.1016/j.eswa.2021.115561.

[5] H. Zhang, Z. Zang, H. Zhu, M. I. Uddin, and M. A. Amin, "Big data-assisted social media analytics for business model for business decision making system competitive analysis," Inf Process Manag, vol. 59, no. 1, 2022, doi: https://www.sciencedirect.com/science/article/pii/S0306457321002430.

[6] T. Issa and P. Isaias, "Usability and Human–Computer Interaction (HCI)," in Sustainable Design, London: Springer London, 2022, pp. 23–40. doi: 10.1007/978-1-4471-7513-1_2.

[7] A. Adel, "Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas," Journal of Cloud Computing, vol. 11, no. 1, 2022, doi: 10.1186/s13677-022-00314-5.

[8] J. L. Hopkins, "An investigation into emerging industry 4.0 technologies as drivers of supply chain innovation in Australia," Comput Ind, vol. 125, no. 103323, 2021, doi: https://doi.org/10.1016/j.compind.2020.103323.

[9] S. S. Alrumiah and M. Hadwan, "Implementing big data analytics in e-commerce: Vendor and customer view," IEEE Access, vol. 9, pp. 37281–37286, 2021, doi: 10.1109/ACCESS.2021.3063615.

[10] L. Li and J. Zhang, "Research and Analysis of an Enterprise E-Commerce Marketing System Under the Big Data Environment," Journal of Organizational and End User Computing, vol. 33, no. 6, pp. 1–19, 2021, doi: 10.4018/joeuc.20211101.oa15.

[11] A. Kamalaldin, D. Sjödin, D. Hullova, and V. Parida, "Configuring ecosystem strategies for digitally enabled process innovation: A framework for equipment suppliers in the process industries," Technovation, vol. 105, no. December 2019, 2021, doi: 10.1016/j.technovation.2021.102250.

[12] C. Janiesch, B. Dinter, P. Mikalef, and O. Tona, "Business analytics and big data research in information systems," Journal of Business Analytics, vol. 5, no. 1, pp. 1–7, 2022, doi: 10.1080/2573234X.2022.2069426.

[13] V. G. Goulart, L. B. Liboni, and L. O. Cezarino, "Balancing skills in the digital transformation era: The future of jobs and the role of higher education," Industry and Higher Education, vol. 36, no. 2, 2021, doi: https://doi.org/10.1177/095042222110297.

[14] O. Cico, L. Jaccheri, A. Nguyen-Duc, and H. Zhang, "Exploring the intersection between software industry and Software Engineering education - A systematic mapping of Software Engineering Trends," Journal of Systems and Software, vol. 172, 2021, doi: 10.1016/j.jss.2020.110736.

[15] J. Miranda et al., "The core components of education 4.0 in higher education: Three case studies in engineering education," Computers and Electrical Engineering, vol. 93, no. June, 2021, doi: 10.1016/j.compeleceng.2021.107278.

[16] "U.S Bureau of Labor Statistics." Accessed: Jan. 13, 2022. [Online]. Available: https://www.bls.gov/ooh/computer-and-information-technology/home.htm

[17] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for Big Data professions: A systematic classification of job roles and required skill sets," Inf Process Manag, vol. 54, no. 5, pp. 807–817, Sep. 2018, doi: 10.1016/j.ipm.2017.05.004.

[18] K. Mahmood, G. Rasool, F. Sabir, and A. Athar, "An Empirical Study of Web Services Topics in Web Developer Discussions on Stack Overflow," IEEE Access, vol. 11, no. February, pp. 9627–9655, 2023, doi: 10.1109/ACCESS.2023.3238813.

[19] H. A. Goh, C. K. Ho, and F. S. Abas, "Front-end deep learning web apps development and deployment: a review," Applied Intelligence, vol. 53, no. 12, pp. 15923–15945, 2023, doi: 10.1007/s10489-022-04278-6.

[20] N. Jayachandran, A. Abdrabou, N. Yamane, and A. Al-Dulaimi, "A Platform for Integrating Internet of Things, Machine Learning, and Big Data Practicum in Electrical Engineering Curricula," Computers, vol. 13, no. 8, p. 198, Aug. 2024, doi: 10.3390/computers13080198.

[21] G. Li, C. Yuan, S. Kamarthi, M. Moghaddam, and X. Jin, "Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis," J Manuf Syst, vol. 60, pp. 692–706, Jul. 2021, doi: 10.1016/j.jmsy.2021.07.007.

[22] M. Labbé, M. Landete, and M. Leal, "Dendrograms, minimum spanning trees and feature selection," Eur J Oper Res, vol. 308, no. 2, pp. 555–567, Jul. 2023, doi: 10.1016/j.ejor.2022.11.031.

[23] P. C. Siswipraptini, H. L. H. S. Warnars, A. Ramadhan, and W. Budiharto, "Information Technology Job Profile using Average-Linkage Hierarchical Clustering Analysis," IEEE Access, vol. 11, no. September, pp. 94647–94663, 2023, doi: 10.1109/ACCESS.2023.3311203.

[24] "Glints." Accessed: Jul. 15, 2024. [Online]. Available: https://glints.com/id/en

[25] "Tech in Asia." Accessed: Jul. 15, 2024. [Online]. Available: https://www.techinasia.com/

[26] "LinkedIn." Accessed: Jul. 15, 2024. [Online]. Available: https://www.linkedin.com/

[27] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Anal, vol. 1, no. 1, p. 9, Dec. 2016, doi: 10.1186/s41044-016-0014-0.

[28] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102–4107, Sep. 2017, doi: 10.3923/jeasci.2017.4102.4107.

[29] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," Knowl Based Syst, vol. 98, pp. 1–29, Apr. 2016, doi: 10.1016/j.knosys.2015.12.006.

[30] N. M. Nawi, W. H. Atomi, and M. Z. Rehman, "The Effect of Data Pre-processing on Optimized Training of Artificial Neural Networks," Procedia Technology, vol. 11, pp. 32–39, 2013, doi: 10.1016/j.protcy.2013.12.159.

[31] R. Friedman, "Tokenization in the Theory of Knowledge," Encyclopedia, vol. 3, no. 1, pp. 380–386, Mar. 2023, doi: 10.3390/encyclopedia3010024.

[32] M. Kashina, I. D. Lenivtceva, and G. D. Kopanitsa, "Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification," Procedia Comput Sci, vol. 178, pp. 284–290, 2020, doi: 10.1016/j.procs.2020.11.030.

[33] E. Elakiya and N. Rajkumar, "Designing preprocessing framework (ERT) for text mining application," in 2017 International Conference on IoT and Application (ICIOT), IEEE, May 2017, pp. 1–8. doi: 10.1109/ICIOTA.2017.8073613.

[34] S. Ahmad and R. Varma, "Information extraction from text messages using data mining techniques," Malaya Journal of Matematik, vol. S, no. 1, pp. 26–29, Jan. 2018, doi: 10.26637/MJM0S01/05.

[35] M. Fachrurrozi, N. Yusliani, and M. M. Agustin, "Identification of Ambiguous Sentence Pattern in Indonesian Using Shift-Reduce Parsing," 2014.

[36] P. EBDEN and R. SPROAT, "The Kestrel TTS text normalization system," Nat Lang Eng, vol. 21, no. 3, pp. 333–353, May 2015, doi: 10.1017/S1351324914000175.

[37] M. Khader, A. Awajan, and G. Al-Naymat, "The Effects of Natural Language Processing on Big Data Analysis: Sentiment Analysis Case Study," in 2018 International Arab Conference on Information Technology (ACIT), IEEE, Nov. 2018, pp. 1–7. doi: 10.1109/ACIT.2018.8672697.

[38] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, Mar. 2020, pp. 466–472. doi: 10.1109/ICACCS48705.2020.9074166.

[39] S. M. Basha and D. S. Rajput, "Evaluating the Impact of Feature Selection on Overall Performance of Sentiment Analysis," in Proceedings of the 2017 International Conference on Information Technology, New York, NY, USA: ACM, Dec. 2017, pp. 96–102. doi: 10.1145/3176653.3176665.

[40] D. R. Rakhimova and A. O. Turganbaeva, "Normalization of Kazakh language words," Scientific and Technical Journal of Information Technologies, Mechanics and Optics, vol. 20, no. 4, pp. 545–551, Aug. 2020, doi: 10.17586/2226-1494-2020-20-4-545-551.

[41] N. Yusliani, R. Primartha, and M. Diana, "Multiprocessing Stemming: A Case Study of Indonesian Stemming," Int J Comput Appl, vol. 182, no. 40, pp. 15–19, Feb. 2019, doi: 10.5120/ijca2019918476.

[42] A. S. Rizki, A. Tjahyanto, and R. Trialih, "Comparison of stemming algorithms on Indonesian text processing," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 17, no. 1, p. 95, Feb. 2019, doi: 10.12928/telkomnika.v17i1.10183.

[43] R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan, and Anderies, "Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity," in 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/ICITDA55840.2022.9971451.

[44] M. Javed and S. Kamal, "Normalization of Unstructured and Informal Text in Sentiment Analysis," International Journal of Advanced Computer Science and Applications, vol. 9, no. 10, 2018, doi: 10.14569/IJACSA.2018.091011.

[45] L. Wang, "Design of Network Public Opinion Monitoring System based on LDA Model," in 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICICACS60521.2024.10498592.

[46] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements," Int J Inf Manage, vol. 50, pp. 416–431, Feb. 2020, doi: 10.1016/j.ijinfomgt.2019.07.014.

[47] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, and F. Amenta, "Text mining with sentiment analysis on seafarers' medical documents," International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100005, Apr. 2021, doi: 10.1016/j.jjimei.2020.100005.

[48] J. E. Montandon, C. Politowski, L. L. Silva, M. T. Valente, F. Petrillo, and Y.-G. Guéhéneuc, "What skills do IT companies look for in new developers? A study with Stack Overflow jobs," Inf Softw Technol, vol. 129, p. 106429, Jan. 2021, doi: 10.1016/j.infsof.2020.106429.

[49] L. T. Khrais, "Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce," Future Internet, vol. 12, no. 12, p. 226, Dec. 2020, doi: 10.3390/fi12120226.

[50] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro, "Sentiment analysis for mining texts and social networks data: Methods and tools," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 1, Jan. 2020, doi: 10.1002/widm.1333.

[51] H. Ren, Y. Liu, G. Naren, and J. Lu, "The impact of multidirectional text typography on text readability in word clouds," Displays, vol. 83, p. 102724, Jul. 2024, doi: 10.1016/j.displa.2024.102724.

[52] J. Lamri and T. Lubart, "Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach," J Intell, vol. 11, no. 6, p. 107, Jun. 2023, doi: 10.3390/jintelligence11060107.

[53] M. Hirudayaraj, R. Baker, F. Baker, and M. Eastman, "Soft Skills for Entry-Level Engineers: What Employers Want," Educ Sci (Basel), vol. 11, no. 10, p. 641, Oct. 2021, doi: 10.3390/educsci11100641.

[54] F. Gurcan and N. E. Cagiltay, "Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling," IEEE Access, vol. 7, pp. 82541–82552, 2019, doi: 10.1109/ACCESS.2019.2924075.

[55] K. Bajaj, K. Pattabiraman, and A. Mesbah, "Mining questions asked by web developers," in Proceedings of the 11th Working Conference on Mining Software Repositories, New York, NY, USA: ACM, May 2014, pp. 112–121. doi: 10.1145/2597073.2597083.