Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Exploring Job Vacancy Topics and Trends in Indonesia Using Latent Dirichlet Allocation (LDA) and Exploratory Data Analysis (EDA)

Yoga Prasetyo Wibowo

Universitas Bina Sarana Informatika, Depok, Indonesia
yogaprasetyo9h@gmail.com

Abstract

The Indonesian labor market is rapidly evolving, creating a need for effective analysis to uncover patterns and trends in job vacancies, especially in the digital era where information is abundant and diverse. This study aims to identify dominant topics and vacancy trends to provide insights for job seekers, companies, and stakeholders in developing targeted recruitment strategies. Data were collected through web scraping from the Jobstreet website, covering company, location, position, classification, subclassification, salary, and posting time. Two methods were applied: Exploratory Data Analysis (EDA) to describe vacancy distributions, and Latent Dirichlet Allocation (LDA) to extract thematic structures from text data. EDA results show that Sales Executive is the most demanded position, Greater Jakarta is the location with the highest vacancies, and the most common salary range is IDR 4–5 million. LDA, validated using the Coherence Score, identified two distinct topics: (1) Sales, Marketing, Supply Chain, and Retail, and (2) Customer Service and Administration.. This research contributes by combining descriptive and text-mining techniques on the latest job vacancy dataset obtained directly through web scraping, producing an analysis that not only highlights vacancy distribution but also reveals latent topics reflecting Indonesian labor market trends.

Keywords: *Exploratory Data Analysis, Latent Dirichlet Allocation, Job Vacancies, Jobstreet, Job Trends*

1. Introduction

The development of information technology has had a significant impact on the world of work, particularly in the recruitment process. Online job portals like Jobstreet are now the primary link between companies and job seekers. Thousands of job advertisements are published daily, containing crucial information about positions, companies, locations, and salary ranges. This information is not only valuable for job seekers but also holds significant potential for analysis to understand labor market patterns and trends. Unfortunately, much of this data remains underutilized due to its unstructured format and massive volume.

A number of studies have attempted to analyze job vacancy data using an analytical approach. For example, applying the Hierarchical Agglomerative Clustering method to group job vacancy data based on education and work experience. The results show that the majority of vacancies require a bachelor's degree with at least one year of experience, with Information Systems majors being the most popular, and the majority of companies are in the manufacturing, finance, and software sectors in South Jakarta [1]. The effectiveness of Latent Dirichlet Allocation (LDA) in discovering hidden topics from unstructured data, such as those used in digital tourism data, is demonstrated. These findings confirm that topic-based analysis approaches can be widely used, including in the context of the job market [2].

However, research on exploring job vacancy trends in Indonesia that combines exploratory analysis and topic modeling is still limited. Therefore, this study applies Exploratory Data Analysis (EDA) to explore job vacancy distribution patterns based on location, company, posting time, and salary range. It also uses LDA to identify key topics from job advertisement texts.

This study aims to provide a comprehensive overview of job market trends in Indonesia, particularly in the Greater Jakarta area. The results are expected to benefit job seekers in developing career strategies, companies in designing more targeted recruitment strategies, and serve as an academic contribution to the application of data analysis methods in the employment sector.

2. Literature Review

2.1. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on how humans and computers can interact using natural language [3]. Its main concern is how machines understand human language to interact with each other. With NLP, computers can learn and understand human language, allowing them to communicate with humans [4]. In its application, NLP uses various algorithms and statistical methods to analyze the structure, meaning, and context of human language. NLP technology is widely used in various applications, such as search engines, chatbots, automatic translation systems, text processing, and sentiment analysis.

2.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a method for identifying patterns, structures, and relationships among data through graphical visualization, making it easier for researchers to understand the data. Most EDA techniques are graphical in nature, supported by several quantitative methods. [5]. The main goal of Exploratory Data Analysis (EDA) is to understand the data in depth before performing more in-depth analysis or building predictive models [6]. EDA helps us uncover data structure, identify patterns, anomalies, and relationships between variables. This process involves data visualization, descriptive statistical calculations, and other techniques to gain insight into the data.

2.3. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative statistical model used to discover hidden topic patterns in a collection of documents. This model assumes that each document is a mixture of several topics, and each topic is represented by a specific distribution of words[7]. In its application, LDA forms a system for determining several topic groups from data or information consisting of words [8]. This study analyzes the number of topics by conducting an evaluation using the coherence value. The coherence value can measure the relationship between words in topic modeling, the best topic evaluation is assessed based on the highest coherence value obtained [9].

2.4. Web Scraping

Web scraping is a technique for extracting data from websites and storing it in a local file format or database. As a subset of data mining, this technique involves various fields such as machine learning and statistics [10]. The advantage of using web scraping is that it is quite time efficient, allows for regular data retrieval within a short period of time, and can obtain a greater amount of information [11]. However, there are some drawbacks, such as copyright infringement, reliance on HTML structure, and ethical and legal challenges [12]. In the context of this research, web scraping is carried out to retrieve information from the Jobstreet website, which retrieves data such as position, location, field of work, time of posting job vacancies.

3. Research Method

The steps in this description follow a flowchart that describes each stage performed during the study. Figure 1 below is:

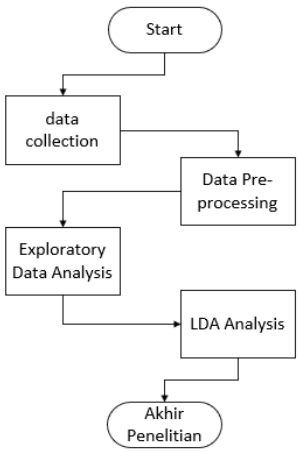

Fig. 1: Research Process and Steps

3.1. Data Collection

Data collection from the Jobstreet job vacancy site on June 21, 2025 using web scraping techniques. The URL used for the scraping process is https://id.jobstreet.com/id/jobs?page=&sortmode=ListedDate. The data collected includes job position, field of work, location, company name. The search results will be web pages in HTML and/or XML format. The retrieved web pages will then be parsed using the BeautifulSoup library [13]. Scraping is used to retrieve data incrementally per page. By using While True, scraping is performed page by page, starting with the first page. Then, at each iteration, a dynamic URL is generated based on the page number and an HTTP request is

sent using the requests library, including a User-Agent header to ensure the request is treated as coming from a regular browser and is not blocked by the server.

	Posisi	Perusahaan	Lokasi	Job Sub Klasifikasi	Job Klasifikasi	Gaji	Waktu
0	Sales Canvasser (All Level)	Pengiklan Anonim	Surakarta	Asisten Ritel	(Ritel & Produk Konsumen)	Gaji tidak disebutkan	Baru saja
1	Assistant Manager, Creative & Brand Assurance	Pacific Licensing Studio Pte Ltd	Jakarta Raya	Desain Grafis	(Desain & Arsitektur)	Gaji tidak disebutkan	Baru saja
2	SCM Supervisor	PT Nippon Indosari Corpindo Tbk.	Surabaya	Pemimpin Tim/Supervisor	(Manufaktur, Transportasi & Logistik)	Gaji tidak disebutkan	Baru saja
3	INTERNAL AUDIT STAFF	PT. Dwidaya World Wide	Jakarta Raya	Audit – Internal	(Akuntansi)	Gaji tidak disebutkan	Baru saja
4	PURCHASING	PT. Bentang Persada Internusa	Surabaya	Pembelian, Pengadaan & Inventaris	(Manufaktur, Transportasi & Logistik)	Gaji tidak disebutkan	1 menit yang lalu

Fig. 2: Sample data that was successfully scraped

3.2. Data Pre-processing

Pre-processing is to clean and prepare data so that it is easier to analyze and produces more accurate results [14]. In the pre-processing process there are two stages, namely assessing data and cleaning data. In the data assessment stage, an evaluation is carried out on the quality of the data that has been collected to identify problems such as missing data, duplication, and format inconsistencies. And Data cleaning is carried out to clean and replace data types to facilitate the analysis process.

3.3. Exploratory Data Analysis Method

In the Exploratory Data Analysis (EDA) stage, this study focused on extracting initial information from the job vacancy dataset obtained through web scraping from the Jobstreet portal. The analysis was conducted by identifying the distribution of vacancies based on key categories, such as location, company, job type, and salary range. Furthermore, data visualization was performed to display trends in the number of job vacancies per specific time period. The purpose of this EDA was to understand the basic characteristics of the data, discover distribution patterns, and identify underlying trends in the Indonesian job market, thus providing a foundation for further analysis using the Latent Dirichlet Allocation (LDA) method.

3.4. LDA Analysis

In the Latent Dirichlet Allocation (LDA) stage, this study uses a topic modeling method to identify hidden topics from the obtained job vacancy texts. This process begins with text data preprocessing, including special character cleaning, word normalization, and stopword removal to ensure more accurate analysis results. Next, LDA is applied to text columns containing job position descriptions, classifications, and sub-classifications, with the aim of finding groups of words that frequently appear together and form specific topics.

4. Results And Discussion

This section presents the implementation of job vacancy data analysis results obtained through web scraping from the Jobstreet portal. The analysis begins with an Exploratory Data Analysis (EDA) stage to explore job vacancy distribution patterns based on various categories, followed by the application of the Latent Dirichlet Allocation (LDA) method to identify key topics emerging from job vacancy texts. The results obtained at each stage are then interpreted and linked to job market trends in Indonesia.

4.1. Research Results

4.1.1. Exploratory Data Analysis

The application of the Exploratory Data Analysis (EDA) method to identify job vacancy distribution patterns based on the main categories contained in the dataset yields some insights into job vacancy trends in 2025.

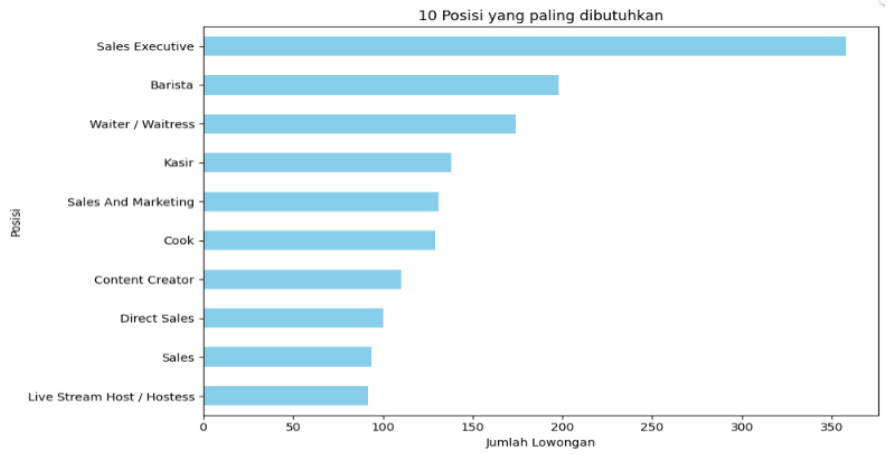

Fig. 3: Most needed positions

Figure 3 shows a visualization of 10 positions most in demand by companies. Sales Executive is the most sought-after position. Available jobs are predominantly in the sales field, including sales assistant positions.

Fig. 4: Most job vacancy locations

The 10 locations with the most job openings. Many companies tag their locations as "Greater Jakarta," making it the location with the most job openings. Next, South Jakarta and North Jakarta. Outside Greater Jakarta, Surabaya ranks fourth with the most job openings, followed by Bandung in West Java, Yogyakarta in Central Java, and Denpasar in Bali. These are the cities in Indonesia with the most job openings, dominated by Jakarta.

Table 2: Top 5 Distribution of Vacancies by Location

No	Location	Number of Vacancies
1	Jakarta Raya	2494
2	Surabaya	1376
3	Tanggerang	900
4	Bandung	764
5	Denpasar	679

Fig. 5: Companies with the most vacancies

Figure 5 shows Companies with numerous job openings. More than 1,000 companies have been identified by hiding their company names under the name "Anonymous Advertiser." This indicates that many job openings are published without clearly identifying the company. Many vacancies are available at outsourcing service providers, which dominate job advertisements.

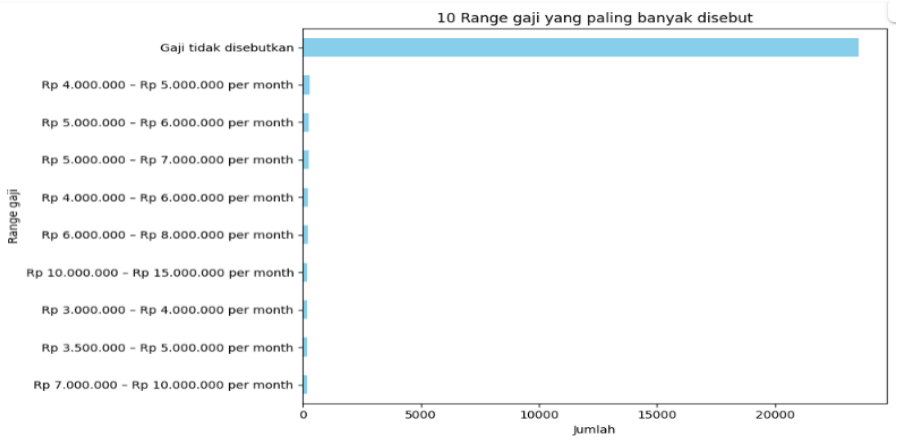

Fig. 6: Range salary

Figure 6 displays the visualization of salary ranges. It can be seen that many companies don't display salaries in job advertisements. Furthermore, the 4-5 million salary range ranks second among companies offering salaries. The smallest salary range is 3-4 million, and the largest salary range is 10-15 million, which ranks third. These findings indicate that many companies are not transparent about salary disclosures.

Table 3: Top 5 salary ranges

No	Location	Number of Vacancies
1	3 – 4 juta	288
2	5 – 6 juta	270
3	5 – 7 juta	264
4	4 – 6 juta	225
5	6 – 8 juta	211

4.1.2. Latent Dirichlet Allocation

Before entering the Latent Dirichlet Allocation model, the position, job classification, and job subclassification columns will be combined to enrich the topic. Then, preprocessing will be carried out to remove punctuation, symbols, and tokenize sentences into words. Next, common words that do not have a significant contribution to the analysis, also known as stopwords, are removed by referring to the Indonesian stopword list from the NLTK library.

```
from gensim.models import CoherenceModel
from gensim.models.ldamodel import LdaModel

coherence_scores = []
for k in range(2, 11):
    model = LdaModel(corpus=corpus, id2word=dictionary, num_topics=k, passes=10, random_state=42)
    coherence = CoherenceModel(model=model, texts=df_jobs['tokens'], dictionary=dictionary, coherence='c_v')
    score = coherence.get_coherence()
    coherence_scores.append((k, score))
    print(f'Jumlah topik: {k}, Coherence Score: {score:.4f}')
```

```
Jumlah topik: 2, Coherence Score: 0.5798
Jumlah topik: 3, Coherence Score: 0.5329
Jumlah topik: 4, Coherence Score: 0.5395
Jumlah topik: 5, Coherence Score: 0.5289
Jumlah topik: 6, Coherence Score: 0.5415
Jumlah topik: 7, Coherence Score: 0.5184
Jumlah topik: 8, Coherence Score: 0.5115
Jumlah topik: 9, Coherence Score: 0.5181
Jumlah topik: 10, Coherence Score: 0.5110
```

Fig. 7: Determining the best number of topics

Next, determine the best number of topics in the topic modeling by conducting an evaluation process using the Coherence Score metric. This process is carried out by trying several topics, namely from 2 to 10 topics. The number of topics selected is the one with the highest Coherence Score. As seen in Figure 7, topic 2 has the highest score, so this study will use 2 topics in the LDA model.

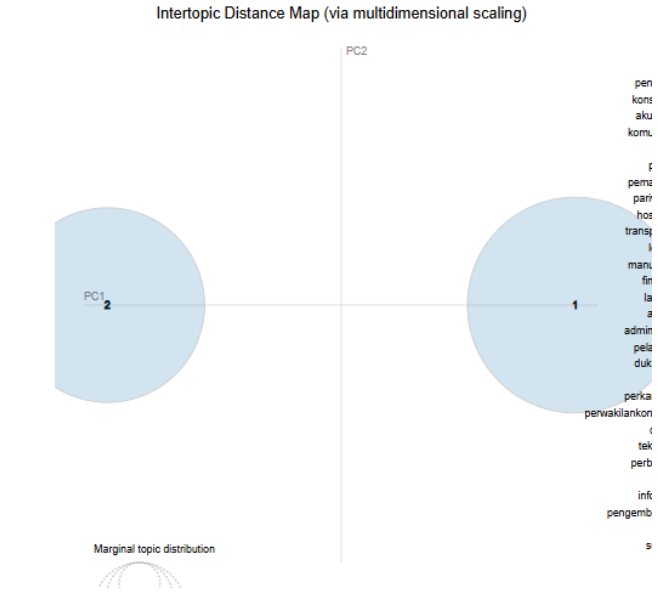

Fig. 8: Intertopic Distance Map Visualization

After the topic has been determined, the next step is to display the results of the topic using the Intertopic Distance Map. In this map, each blue circle represents one topic identified by the LDA model, where the size of the circle reflects how often the topic appears in the entire document. The two topics displayed appear clearly separated and do not overlap, indicating that the model is able to distinguish topics semantically. The greater distance between topics reflects significant differences in the set of words that make them up, so it can be concluded that the resulting topics are quite separate and not confusing.

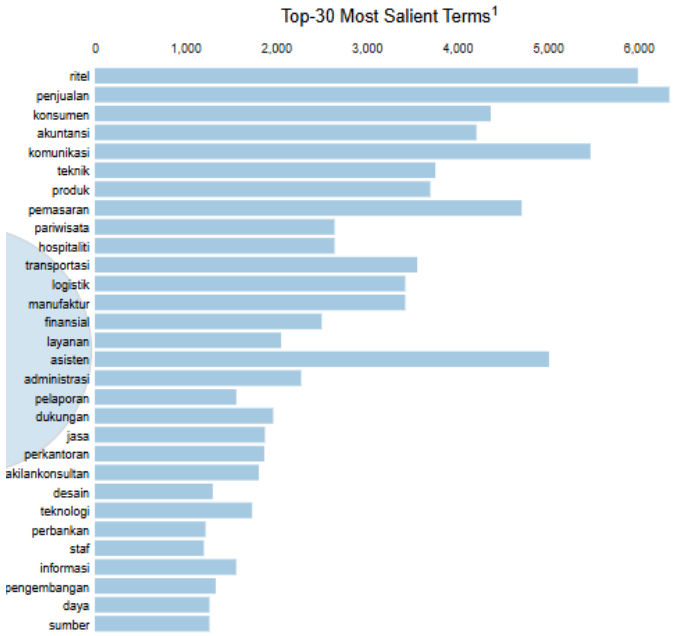

Fig. 9: Salient terms visualization

Figure 9 displays the 30 most salient terms appearing across the dataset. This graph displays the frequency of word occurrences as horizontal bars, with blue bars indicating the frequency of the word across all documents. Words such as sales, marketing, accounting, and communication appear to dominate and have high frequencies, indicating that these terms are key indicators in topic formation.

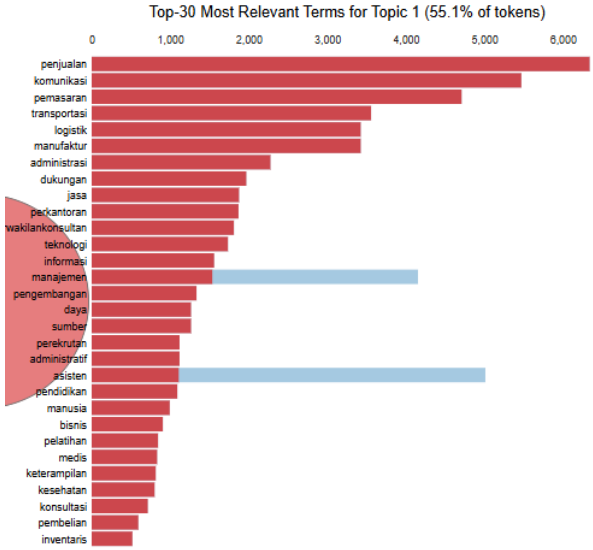

Fig. 10: Salient terms visualization topic 1

The results of topic model 1, which covers approximately 55% of the total tokens in the job posting corpus, include dominant keywords such as sales, communications, marketing, transportation, logistics, manufacturing, and administration. These keywords describe job fields that focus on sales and marketing activities, along with supply chain connections such as transportation and logistics. This topic can be categorized as "Sales, Marketing, and Supply Chain Jobs."

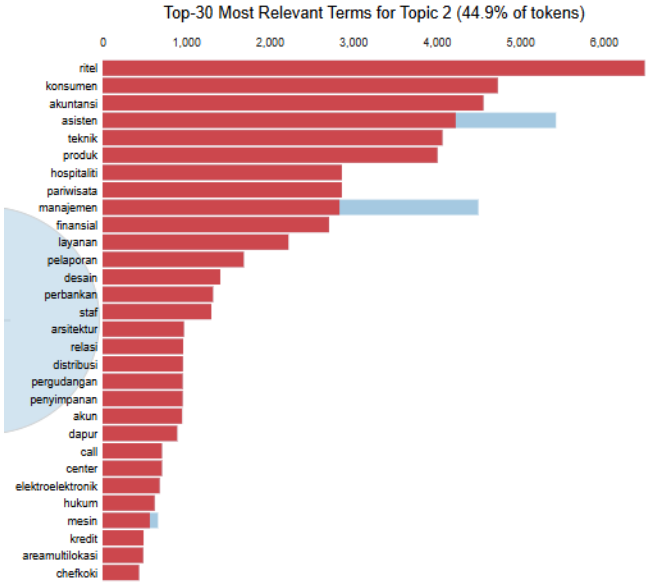

Fig. 11: Salient terms visualization topic 2

Topic 2, which covers approximately 44.9% of the total tokens in the job posting corpus, includes keywords such as retail, consumer, accounting, assistant, engineering, product, hospitality, tourism, and management. These keywords indicate job fields related to consumer services, the retail sector, financial administration, and the hospitality and tourism industry. This topic can be identified as “Retail, Consumer Services, and Administration Jobs.”

4.2. Discussion

In the LDA method, evaluating the optimal number of topics using the Coherence Score yielded the highest value at 2 topics, which were then used for the final modeling. These results indicate that the model can group job vacancy data into two semantically distinct broad themes:

Table 4: LDA modeling topic

Topic	Frequently occurring words	Token distribution
1	sales, communications, marketing, transportation, logistics, manufacturing, and administration	55,1%
2	retail, consumer, accounting, assistant, engineering, product, hospitality, tourism, and management.	44,9%

Topic 1 reflects jobs in Sales, Marketing, and Supply Chain, and topic 2 reflects jobs in Retail, Customer Service, and Administration. The Intertopic Distance Map visualization shows that the two topics are quite far apart, thus concluding that the topics generated by the LDA model are significantly different. This indicates that the LDA method successfully identified hidden thematic patterns in the dataset.

The EDA method also yielded consistent results with the LDA method. The distribution of job titles, locations, companies, and salary ranges showed trends that supported the LDA topic interpretation. For example, the dominance of Sales Executive positions and the Greater Jakarta location aligns with the sales and customer service topics identified in the LDA. These results are expected to provide meaningful and useful information for job seekers in understanding changes in the labor market, as well as for companies in reassessing their recruitment strategies in line with existing trends.

5. Conclusion

This study presents a novel integration of Exploratory Data Analysis (EDA) and Latent Dirichlet Allocation (LDA) to explore job vacancies in Indonesia. The results highlight the dominance of Sales Executive positions, the concentration of opportunities in Greater Jakarta (Jabodetabek), and the prevalence of salaries around IDR 4–5 million. More importantly, LDA uncovers two key topics: (1) Sales, Marketing, Supply Chain, and Retail and (2) Customer Service and Administration. These findings suggest that the Indonesian labor market is heavily driven by sales and service, offering new insights rarely found in previous studies. The contribution of this research lies in how the combination of EDA and LDA can provide a more comprehensive understanding of labor market

6. Suggestions

Suggestions that can be given for further research are as follows:

1. Add analysis variables such as education, work experience, and job description to make the topic more detailed.
2. Combine LDA with other methods, such as clustering or text classification, to increase the validity of the results.
3. Compare with data from other job boards to verify the consistency of trends.

References

[1] 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.

[2] N. C. K.Uray, "Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA)," vol. 13, no. 4, pp. 6075–6086, 2024.

[3] E. S. Eriana and D. A. Zein, "Artificial Intelligence," *Angew. Chemie Int. Ed.*, vol. 6(11), p. 1, 2023.

[4] Y. Waruwu, "Pendidikan Agama Kristen Dalam Era Ai: Menggunakan Kecerdasan Buatan Untuk Personalisasi Pembelajaran Spiritual," *J. Abdiel Khazanah Pemikir. Teol. Pendidik. Agama Kristen dan Musik Gereja*, vol. 8, no. 2, pp. 151–165, 2024, doi: 10.37368/ja.v8i2.786.

[5] D. Leni, F. Earnestly, R. Sumiati, A. Adriansyah, and Y. P. Kusuma, "Evaluasi sifat mekanik baja paduan rendah bedasarkan komposisi kimia dan suhu perlakuan panas menggunakan teknik exploratory data analysis (EDA)," *Din. Tek. Mesin*, vol. 13, no. 1, p. 74, 2023, doi: 10.29303/dtm.v13i1.624.

[6] F. Alfiah *et al.*, *Pemodelan Dan Visualisasi Data*, no. June. 2025.

[7] Angga Reni Dwi Astuti and N. Cahyono, "Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 326–334, 2023, doi: 10.33022/ijcs.v12i1.3155.

[8] C. Natalia, F. Suprata, F. P. S. Surbakti, and S. Clarence, "Penentuan Standar Spesifikasi Kerja di Café Berdasarkan Big Data dengan Metode LDA dan AHP," *J. Rekayasa Sist. Ind.*, vol. 10, no. 2, pp. 211–226, 2021, doi: 10.26593/jrsi.v10i2.5228.211-226.

[9] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Topic Modeling on Online News Media Using Latent Diriclet Allocation (Case Study Somethinc Brand)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 481–489, 2024.

[10] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 4, pp. 377–383, 2023, doi: 10.47233/jteksis.v5i4.913.

[11] Ivana Elfirdaus and Eka Dyar Wahyuni, "Implementasi Web Scraping Untuk Pengambilan Data Rekomendasi Film Pada Imdb," *Pros. Semin. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 327–333, 2023, doi: 10.33005/sitasi.v3i1.647.

[12] D. Chrisinta and J. E. Simarmata, "Eksplorasi Teknik Web Scraping pada Data Mining: Pendekatan Pencarian Data Berbasis Python," *Fakt. Exacta*, vol. 17, no. 1, pp. 58–68, 2024, doi: 10.30998/faktorexacta.v17i1.22393.

[13] N. Fadhilla Rosia Afrianti and A. Badawi, "Web Scraping Senyawa Herbal Di Indonesia Menggunakan Selenium Python," *J. Sci. Soc. Res.*, vol. 4307, no. 4, pp. 1362–1366, 2024, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[14] M. Z. Haq, C. S. Octiva, A. Ayuliana, U. W. Nuryanto, and D. Suryadi, "Algoritma Naïve Bayes untuk Mengidentifikasi Hoaks di Media Sosial," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1079–1084, 2024, doi: 10.33395/jmp.v13i1.13937.