From Words to Wages: Understanding Salary Determinants Through Job Description Text Mining in Indonesia

1st Raja Adi Pranata
Computer Science and Statistics
Universitas Bina Nusantara
Jakarta, Indonesia
raja.pranata@binus.ac.id

2nd Joevan Alezka
Computer Science and Statistics
Universitas Bina Nusantara
Jakarta, Indonesia
joevan.alezka@binus.ac.id

3rd Erna Fransisca Angela Sihotang, S.Stat, M.Kom.
Computer Science and Statistics
Universitas Bina Nusantara
Jakarta, Indonesia
erna.sihotang@binus.ac.id

4th Margaretha Ohyver S.Si., M.Si.
Computer Science and Statistics
Universitas Bina Nusantara
Jakarta, Indonesia
ethaohyver@binus.ac.id

Indonesia's job market is evolving rapidly with digital adoption, shifting skill demands, and uneven wage structures across regions and seniority levels. This study examines how the language of job advertisements relates to posted salaries in Indonesia's data-and-analytics segment. We compile a corpus of analyst-oriented vacancies from JobStreet Indonesia (via Kaggle), cleanse and normalize bilingual text (Bahasa Indonesia and English), and transform job descriptions and listed tools into TF-IDF features. These text features are combined with structured attributes (province, level, experience) and modeled using regularized linear regressions (Ridge and Lasso). Model selection and uncertainty are assessed with 5-fold cross-validation; final performance is reported on a held-out test set using MAE, RMSE, and R^2. Ridge delivers the most stable generalization (test $R^2 \approx 0.21$ with lower MAE/RMSE than the baseline and Lasso), consistent with high-dimensional, correlated TF-IDF inputs. Coefficient interpretation reveals economically meaningful signals: higher salaries are associated with greater years of experience, senior or managerial cues, Jakarta-based locations, and English/business-facing terminology; lower salaries align with entry-level indicators, non-Jakarta locations, and spreadsheet-centric tooling. Residual diagnostics support the adequacy of the log-salary specification while highlighting remaining heteroscedasticity at the upper tail. The contribution is a localized, interpretable framework that translates Indonesian job-ad "words" into measurable wage determinants, informing transparent compensation design, targeted reskilling, and regionally nuanced workforce planning.

Keywords— Ridge, Lasso,Text mining

I. INTRODUCTION

According to Indonesia's Statistics Agency (BPS), in 2023 Indonesia faced a problem in wage setting between men and women. The issue is a significant pay difference of around 16.6 percent, with men earning Rp 20,125 per hour and women earning Rp 16,779 per hour [1]. The underlying cause is that the rapidly changing demand for skills in Indonesia's labor market, driven by digitalization and shifts in the economic structure, has not been matched by corresponding upskilling [2]. In addition, the problem is reinforced by concentration in informal and part-time work. BPS reports that in 2023 the share of informal employment was 64.25% for women and 55.81% for men [1].

In Indonesia's rapidly changing job market, understanding what influences salary levels has become increasingly important for job seekers, employers, and policymakers. While most salary prediction models still depend on structured data such as education, experience, and industry, there is another valuable source that often goes unnoticed which is the text within job descriptions [3]. This study aims to explore that potential by using text mining methods, particularly Term Frequency–Inverse Document Frequency (TF-IDF) [4], along with regression analysis to identify and measure factors that shape salaries in Indonesia.

TF-IDF is a common text-representation scheme that turns job-description words into numeric features by weighting each term by how often it appears in a document relative to how rare it is across all postings, and it has been used effectively for salary prediction from vacancy text[3]. In labor-market analytics, surveys note that classic NLP methods such as TF-IDF still serve as strong baselines within end-to-end pipelines for skills and wage analysis drawn from online vacancies[5]. Recent prototyping endeavors involving job-market data have incorporated TF-IDF with contemporary models to predict wages and categorize occupations. This approach underscores the practicality of TF-IDF in conjunction with more recent embeddings.[6].

The phrase *"From Words to Wages"* from the title reflects the transformation of textual job information (*words*) into measurable insights that explain pay outcomes (*wages*). This highlights how unstructured text data, once processed with text-mining techniques, can directly reveal the determinants of salary levels.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Salary determinants can be understood from various perspectives including education, skills, experience, job role, and especially linguistic expressions found in job postings that may implicitly signal pay levels [3]. This study specifically focuses on *job descriptions* because they contain explicit and implicit indicators of required skills, experience, and responsibilities that employers associate with certain wage ranges.

Job descriptions can be used to understand salary determinants through text-mining methods, particularly TF-IDF and regression analysis. This study proposes a novel approach that integrates TF-IDF-based features with regression models to predict salaries and identify specific words and phrases in job postings that may influence pay levels [7]. The significance of these findings lies in the paucity of literature examining the direct correlation between textural content in job descriptions and salary variations, particularly within Indonesia's labor market. Consequently, the insights obtained in this study are regarded as findings because they augment existing research by unveiling new empirical evidence connecting language features in postings with actual pay outcomes.

The objective of this study is to explore that potential by employing text mining methods, particularly Term Frequency–Inverse Document Frequency (TF-IDF) [4], in conjunction with regression analysis to identify and measure factors that shape salaries in Indonesia.

In short, this paper contributes to both labor economics and data science by showing how unstructured text data can be used for salary prediction in an Indonesian context. The goal is to improve transparency around salary determination and to support more efficient workforce planning and job matching in the country.

II. LITERATURE REVIEW

This Research Using Paper Internationally indexed from 2020-2025

Recent empirical work shows that online job advertisements can reveal salary determinants once text features are extracted and linked to observed or predicted pay outcomes [8]. Comparative evaluations of skill-extraction methods report that the choice of taxonomy or data-driven categorization materially changes the share of wage variation a model can explain [9]. A broad survey of labor-market analytics documents pipelines for turning vacancy text into features and positions TF-IDF as a competitive baseline despite the rise of deep embeddings [10]. Previous studies have shown that TF-IDF works well for turning text into useful numerical features [5], and that regression models can effectively predict salaries.

However, most of those studies concentrate on Western job markets or use general datasets that don’t reflect Indonesia’s unique economic and cultural context [8]. Our research fills that gap by focusing specifically on the Indonesian market, offering insights that are more relevant and localized [8].

Language in postings carries more than role labels because co-mentioned skills and phrasing correlate with wage premia and hiring outcomes in multiple markets [11]. Text rewriting and bias-aware preprocessing can improve the reliability of downstream salary modeling by reducing gendered wording that is correlated with pay gaps [12]. Classical sparse representations such as TF-IDF often perform strongly when paired with linear or regularized regressions for salary prediction tasks, particularly when feature engineering follows robust cleaning and normalization steps [3]. Field evidence on posted-wage experiments underscores why models trained on ads with explicit salary ranges tend to capture stronger signals than models trained on postings without disclosure [13]. Clustering and NLP over national-scale corpora recover coherent skill profiles across sectors, which align with observed wage differentiation and can be fed as structured features into regression models [14] Interactive visual analytics systems built on vacancy text demonstrate how multi-level skill representations can expose wage-relevant patterns and support feature exploration before model fitting [15]. Network-based studies of co-occurring skills show that clusters differ in mean predicted salaries, reinforcing the idea that combinations of skills carry distinct pay signals beyond single keywords [6].

Compared with Pisolla et al. (2025)[3].The present study demonstrates that a lightweight TF-IDF plus ensemble-learning stack can raise predictive accuracy on heterogeneous job-ad corpora. This advancement of the agenda is achieved through the study's four-dimensional approach: the localization of the problem to Indonesia's institutional and linguistic context, the targeting of continuous wage estimation rather than generic classification, the prioritization of policy-grade interpretability via regularized linear models (Ridge/Lasso) and feature attribution, and the explicit linking of findings to regional wage structures and reskilling priorities. Methodologically, we retain TF-IDF for transparency but pair it with cross-validated regularization and out-of-time evaluation to stabilize coefficients under multicollinearity and to surface term- and skill-level contributions to pay. We acknowledge the limitations that are common to vacancy-text research, including selection bias from postings that disclose salaries, platform-specific sampling, lexical constraints of TF-IDF in capturing deeper semantics, multilingual normalization challenges, and sectoral generalizability. This research proposes future extensions with richer embeddings, multi-portal data, and robustness checks across provinces and industries. In summary, while prior studies have established the technical feasibility and accuracy improvements, the present study offers a localized, interpretable, and policy-relevant salary modeling framework. This framework quantifies the manner in which Indonesia-specific job-ad language translates into measurable wage determinants. Furthermore, it produces actionable evidence for employers and policymakers.

III. METHODOLOGY

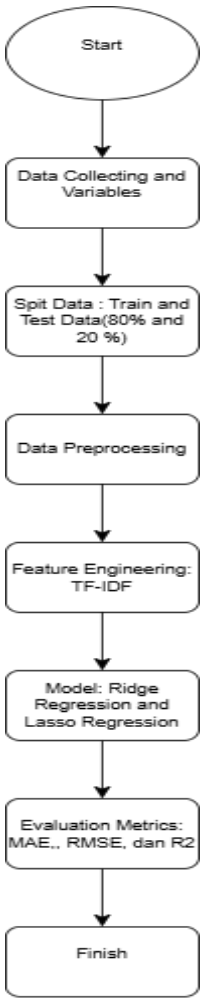


FIGURE 1 METHODOLOGY DIAGRAM

1. Data Collecting and Variables

The dataset used in this research was collected from JobStreet Indonesia, focusing specifically on job advertisements within the data and analytics category. From the complete dataset, only listings that explicitly mentioned "analyst" roles such as Data Analyst, Business Analyst, Financial Analyst, and Risk Analyst were included to ensure relevance to analytical professions. This filtering process resulted in a curated collection of analyst-oriented job postings, providing a clear representation of Indonesia's growing analytical workforce segment.

In addition, the dataset was sourced from a publicly available Kaggle repository titled *"Indonesian Data and Analytics Jobs in JobStreet"* (https://www.kaggle.com/datasets/raflirizkya/indonesian-data-and-analytics-jobs-in-jobstreet/data). The dataset was retrieved from this open-source platform to ensure data accessibility, transparency, and reproducibility of the research findings.

Geographically, this dataset spans 18 out of 38 Indonesian provinces, with the majority of postings concentrated in DKI Jakarta, West Java, East Java, and Bali, which are recognized as the main economic and technological hubs of the country. This limited yet diverse geographical coverage reflects the uneven distribution of data-related job opportunities, where urban centers dominate the recruitment landscape.

Each record in the dataset includes several key variables extracted from the job advertisements: job title, location, company name, job description text, and offered salary. These features were chosen to capture both structured and unstructured aspects of the postings, allowing text-based information (such as job descriptions) to be transformed into measurable predictors through text mining techniques. Prior research emphasizes that combining textual and structured job attributes can significantly improve salary estimation and labor market analysis accuracy [15].

The data collection approach aligns with the practices outlined in recent studies on job market analytics, which recommend using publicly available online postings as proxies for real labor trends while applying rigorous preprocessing to ensure data quality [15]. Moreover, focusing on analyst-specific roles allows the study to contribute localized insights into Indonesia's data economy, complementing global findings on skill demand and wage structure across analytical occupations [15].

2. Split Data

Before conducting model training, the dataset was divided into distinct subsets to ensure objective performance evaluation and reduce the risk of overfitting. The available records were first screened to remove incomplete or inconsistent entries particularly those missing salary information since accurate compensation values are critical for regression-based modeling. After cleaning, the remaining data were split into training (80%) and testing (20%) sets using a randomized stratified approach based on province representation. This strategy preserved the regional distribution of job postings across the 18 provinces, maintaining demographic balance between the two subsets.

The decision to use an 80% and 20% split ratio follows common practices in job advertisement analytics, where sample diversity must be retained while allowing sufficient data for training robust predictive models [12]. Stratification by province was particularly important in this study, as wage structures vary significantly across different regions in Indonesia. Keeping the distribution consistent between training and testing sets ensured that model evaluation reflected realistic geographic variation rather than sampling bias.

Additionally, prior to splitting, all categorical and textual fields such as job title, company, and description were validated for encoding consistency. This step prevented information leakage, ensuring that preprocessing transformations applied during model training (for example, TF–IDF vectorization) were independently re-fitted to the training set and only applied to the test set afterward. This methodological safeguard aligns with best practices in text-based labor market modeling, where reusing fitted transformations can unintentionally bias model performance metrics [16].

By structuring the data-splitting process carefully, this research guarantees that subsequent analyses such as regression modeling and evaluation are grounded on statistically sound partitions that reflect Indonesia's analyst job market fairly and transparently.

3. Data Preprocessing

In preprocessing, cleansing is the first pass that removes noise and fixes obvious defects in the raw postings: lowercasing, stripping HTML tags, punctuation and special characters, collapsing whitespace, and dropping boilerplate; we also delete exact duplicates, handle missing values explicitly (e.g., impute or exclude null salary/education/experience), and standardize obvious typos. Conceptually, this corresponds to the "cleaning, normalization, and deduplication" stage in an end-to-end online-vacancy pipeline that prepares job ads for downstream analysis, with explicit handling of missing data before feature extraction [17].

After the dataset is clean, preprocessing transforms content into learning to ready representations: encode and normalize structured fields (e.g., province, experience), detect language and normalize entities (titles, company, location), then tokenize, remove stopwords, and stem/lemmatize the Indonesian/English text before feature engineering with TF-IDF for modeling; this follows the "clean, normalize entities, extract information" flow for turning raw job postings into structured indicators [17]. In our pipeline, TF-IDF is the bridge from text to numeric features used by the regression models (Ridge/Lasso).

4. Feature Engineering

Feature engineering plays a crucial role in transforming unstructured job description text into meaningful numerical representations that can be analyzed by regression models. In this study, the textual content from the column "deskripsi_lengkap" and skill-related attributes from "tools" were processed using the Term Frequency–Inverse Document Frequency (TF–IDF) method. This approach quantifies the importance of a word within a document relative to its frequency across all job postings, enabling the model to capture distinctive linguistic cues that correlate with salary variation.

Unlike traditional word-count or one-hot encoding, TF–IDF emphasizes unique terms such as *"machine learning," "senior,"* or *"SQL,"* which frequently appear in higher-paying job descriptions but not across all postings. By focusing on relative term weight rather than raw frequency, TF–IDF allows the regression model to learn subtle associations between specialized language and compensation levels. This technique has been widely recognized for its interpretability and efficiency in analyzing job advertisement text, particularly when the dataset includes multilingual or domain-specific terminology [14]. The text-mining pipeline began with tokenization and normalization, followed by bilingual stopword removal (Bahasa Indonesia and English) and stemming using the Sastrawi library. After preprocessing, the cleaned tokens were converted into TF–IDF n-gram vectors (n = 1–2) to preserve both individual keywords and short phrases that often describe responsibilities or tools, such as *"data analysis"* or *"Python programming."* To reduce dimensionality and computational complexity, terms with extremely low document frequency were excluded, and the top 5,000 most informative features were retained.

In parallel, categorical variables such as job title, level, education, and province were one-hot encoded to ensure consistent numerical representation. Numeric features like years of experience were standardized to prevent scale dominance during regression. All structured and unstructured variables were then concatenated into a unified feature matrix, which became the input for the Ridge and Lasso regression models in the next stage.

This feature-engineering design allows the model not only to learn from conventional structured information but also to infer salary determinants directly from text semantics bridging linguistic information with quantitative wage analysis. Studies in labor-market analytics emphasize that such hybrid approaches significantly improve the explainability of salary prediction models by integrating both textual and non-textual attributes [15].

5. Regression Model

5.1. Ridge Regression

Ridge regression is a penalized linear estimator that augments least squares with an L2 penalty to shrink coefficients, thereby reducing variance, mitigating multicollinearity, and improving out-of-sample stability. This study, where Indonesian job-ad text is vectorized into a high-dimensional, sparse TF-IDF design matrix with correlated terms, ridge provides a robust first-line model that preserves signals while controlling coefficient magnitudes, with the regularization strength selected via cross-validation and features standardized, consistent with best practice vacancy to intelligence pipelines for turning online postings into structured indicators for downstream analytics[17]. It is calculated using the formula:

$$\hat{\beta}^{ridge} = arg\frac{min}{\beta}\{\| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2\}$$

FIGURE 2 RIDGE REGRESSION MODEL

5.2. Lasso Regression

Lasso regression is a penalized least-squares estimator with an L1 regularizer that induces sparsity by driving small coefficients exactly to zero, thereby performing embedded feature selection and improving generalization in high-dimensional TF-IDF representations [5].

Ridge regression, with its L2 penalty that shrinks groups of correlated coefficients without setting them to zero, is adopted in this study to stabilize estimates under multicollinearity in Indonesian job-ad TF-IDF features, control variance, and enhance out-of-sample salary prediction while preserving interpretable coefficient patterns[5]. It is calculated using the formula:

$$\hat{\beta}^{lasso} = arg\frac{min}{\beta}\{\| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \}$$

FIGURE 3 LASSO REGRESSION MODEL

6. Evaluation

Model evaluation is a crucial step to assess how accurately the regression models predict job salary values based on textual and structured features. In this study, three key evaluation metrics were employed: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R^2). These metrics were selected because they provide a comprehensive understanding of both the magnitude and proportion of prediction errors, offering insights into how well the Ridge and Lasso regression models generalize on unseen data. The evaluation was conducted on the test subset to ensure unbiased performance measurement and avoid data leakage from the training phase [15].

6.1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) represents the average magnitude of errors between predicted and actual salary values, without considering their direction. It is calculated using the formula:

$$MAE = \frac{1}{n} \Sigma |y_i - \hat{y}_i|$$

FIGURE 4 MEAN ABSOLUTE ERROR CALCULATE

where $\mathbf{y_i}$ denotes the actual salary, $\hat{\mathbf{y}}_\mathbf{i}$ the predicted salary, and **n** the total number of observations. MAE provides an intuitive measure of prediction accuracy since it expresses the average difference between actual and predicted outcomes in the same unit as the target variable. A lower MAE value indicates better predictive performance and smaller deviations from true salary values [18].

6.2. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) measures the square root of the average squared differences between predicted and actual values, using the formula:

$$\mathbf{RMSE} = \sqrt{(\mathbf{1/n})\ \Sigma\ (y_i - \hat{y}_i)^2}$$

FIGURE 5 ROOT MEAN SQUARED ERROR CALCULATE

Unlike MAE, RMSE penalizes larger errors more heavily due to the squaring process, making it more sensitive to outliers. This metric is particularly useful for detecting models that occasionally make large prediction errors. In salary prediction contexts, a smaller RMSE implies that the model can produce consistent predictions without extreme deviations, demonstrating higher stability and reliability [19].

6.3. R-Squared (R^2)

R-Squared, or the coefficient of determination, quantifies the proportion of variance in the dependent variable (salary) that is explained by the independent variables (features). It uses the formula:

$$\mathbf{R^2 = 1 - [\ \Sigma\ (y_i - \hat{y}_i)^2\ /\ \Sigma\ (y_i - \bar{y})^2\ ]}$$

FIGURE 6 R-SQUARED CALCULATE

where $\bar{\mathbf{y}}$ represents the mean of the actual salary values. An R^2 value closer to 1 indicates that the model explains a larger portion of salary variation, signifying better model fit. However, R^2 alone does not indicate error magnitude, which is why it is interpreted alongside MAE and RMSE for a balanced performance evaluation. This multi-metric assessment provides a more holistic evaluation framework, ensuring that the selected regression models perform well both statistically and practically in predicting salary outcomes [19].

IV. ANALYSIS AND DISCUSSION

1. Overview

This study analyzes online job advertisements for data and analytics roles collected from JobStreet Indonesia via a public Kaggle repository. The postings are filtered to analyst-related titles (e.g., Data Analyst, Business Analyst, Financial Analyst, Risk Analyst) to ensure relevance to the targeted occupation group. The dataset covers multiple Indonesian provinces with a strong concentration in major economic hubs, and the job-description text is bilingual (Bahasa Indonesia and English), reflecting common recruitment practices in Indonesia's tech and analytics labor market. The dataset variables used in the study is available below.

Variables	***Explanation***
Posisi(Position)	The title of the job position being offered (e.g., 'Data Analyst', 'Business Intelligence Manager')
Perusahaan(Conpany)	The name of the company that posted the job vacancy
Kota(City)	The city where the job is located
Provinsi(Province)	The province in Indonesia where the job is located
Gaji(Salary)	The offered monthly salary in Indonesian Rupiah (IDR). Note: This column contains many null values as salary is
Tools	A text field listing the required technical skills, tools, and software (e.g., 'SQL, Python, Tableau')
Pendidikan(education)	The minimum or preferred educational background for the role
Pengalaman(Experience)	The minimum required years of work experience
Deskripsi_lengkap	The full text of the job description
Level	The career level of the position

TABLE 1 DATASET VARIABLES

2. Descriptive Statistics and Salary Distribution

This section describes the distributional characteristics of the disclosed monthly salaries in the dataset. Salary data in online job postings commonly exhibit right-skewness and outliers, which can influence model fitting and evaluation. To improve numerical stability and reduce skewness, the salary variable is also examined in a log-transformed form (log(1 + salary)), which typically produces a more symmetric distribution that is easier to model with linear regression methods.

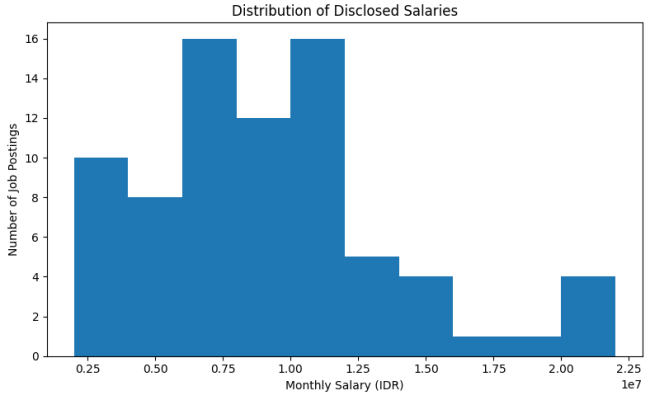


FIGURE 7 DISTRIBUTION OF DISCLOSED SALARIES

The histogram shows a right-skewed distribution of disclosed monthly salaries, with a long thin tail extending to very high values. Most postings sit in a mid-range band roughly around six to twelve million IDR, while there are fewer announcements at the very low end and only a small number of high-salary outliers above about fifteen to twenty-two million IDR. The shape likely reflects mixed seniority tiers in one pool: a lower mode for junior roles and a higher mass for mid-level postings, with a sparse tail of senior or niche roles. This pattern is typical for online vacancies and implies that raw-scale regression will be influenced by a few large salaries.

The visible clustering suggests that language cues in job ads are plausibly tied to these pay bands. In subsequent analyses we will quantify how seniority phrases (for example "senior", "lead", "manager"), tool signals (for example "Python", "SQL", "Tableau", "cloud"), and location terms (city and province) shift the distribution on the log-salary scale. We will also slice the histogram by career level and province to test whether distributional differences align with textual signals, thereby connecting the words in postings to wage outcomes as stated in the study's objective.

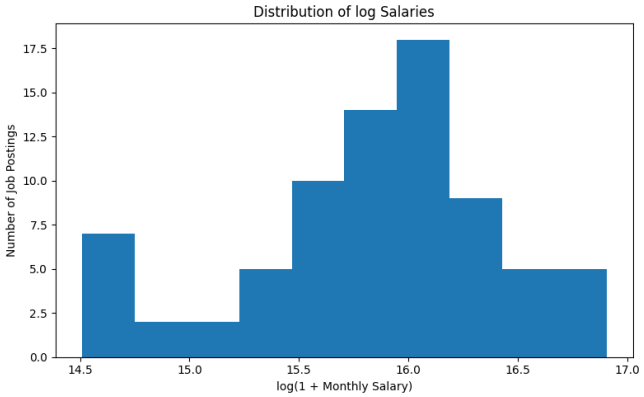


FIGURE 8 DISTRIBUTION OF LOG SALARIES

After the log transform the salary distribution becomes much closer to symmetric and unimodal, centered roughly around log(1 + salary) ≈ 15.8 – 16.1 with a noticeably shorter right tail than on the raw scale. The mass is concentrated in a compact mid-band and there are only a few points at the extremes, which indicates reduced influence of very high salaries on model fitting. This shape is consistent with an approximate log-normal process for posted pay and supports the use of linear models on the transformed target.

For our study, stabilization is important for two reasons. First, Ridge and Lasso assume a linear relationship between features and the target, as well as roughly constant error variance. The log transformation improves both linearity and homoscedasticity. This should result in more reliable coefficients and fairer comparisons across roles, provinces, and seniority levels. Second, interpretability improves because a one-unit change in a textual feature's effect can be interpreted as a multiplicative change in original salaries. Next, we will confirm distributional adequacy using Q-Q plots and residual diagnostics. Then, we will relate shifts in the log-salary distribution to language cues in job ads, such as seniority terms, required tools, and location indicators. This will tie the words in postings to wage outcomes, aligning with the paper's objective.

3. Exploratory Analysis by Seniority Level

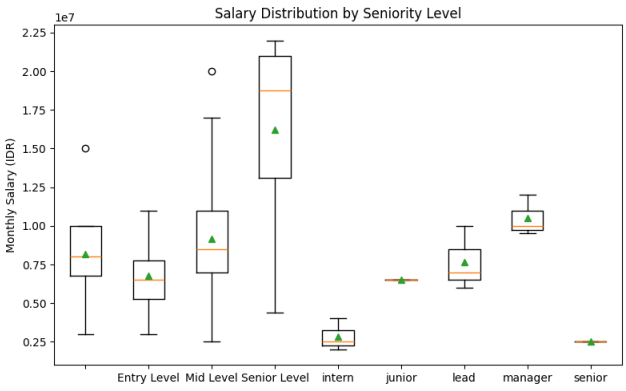


FIGURE 9 . SALARY DISTRIBUTION BY SENIORITY LEVEL

Following the stabilization of the salary distribution through log transformation, the next step examines how compensation varies across inferred seniority levels. Seniority is derived from textual cues in job titles and descriptions, such as *intern*, *junior*, *senior*, *lead*, and *manager*, which are commonly used in online job advertisements to signal experience requirements and responsibility scope.

The boxplot comparison reveals a clear monotonic pattern: median salaries increase steadily with seniority. Entry-level and internship roles exhibit the lowest median log-salaries and relatively narrow interquartile ranges, indicating more standardized compensation. In contrast, senior, lead, and managerial positions show substantially higher medians and wider dispersion, reflecting both higher pay expectations and greater heterogeneity in role scope and organizational context.

Importantly, there remains considerable overlap between adjacent seniority categories, particularly between mid-level and senior roles. This overlap suggests that seniority alone does not fully determine compensation and that additional factors, such as specific technical skills, tools, and location, play a significant role. These observations motivate the inclusion of both structured seniority indicators and rich textual features in the subsequent regression models.

4. Model Validation Using 5-Fold Cross-Validation

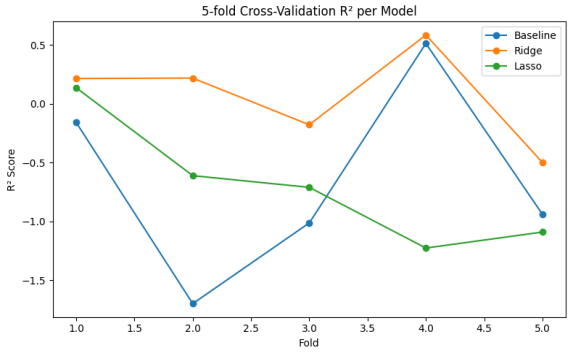


FIGURE 10 5-FOLD CROSS-VALIDATION R^2 PER MODE

We evaluated generalization on the training corpus with a 5-fold cross-validation protocol, reporting the coefficient of determination R^2 on each held-out fold for a baseline linear model (role–province median pay), Ridge, and Lasso. As shown in Figure X, Ridge is consistently the strongest performer: it attains positive R^2 on most folds and peaks at roughly 0.6 on Fold 4, indicating substantive explanatory power relative to a mean predictor. The baseline yields mostly negative R^2 and turns positive only in Fold 4, reflecting poor fit when confronted with unseen postings. Lasso underperforms across folds with predominantly negative R^2, which is consistent with overly aggressive coefficient shrinkage in a high-dimensional TF-IDF space and the relatively small sample size. The dispersion of scores across folds highlights heterogeneity in the postings mix; nevertheless, Ridge remains the dominant and most stable specification. Based on these results, we select Ridge for downstream analysis and interpretation of salary determinants from Indonesian job-ad language.

5. Test-Set Performance Comparison

Model	MAE	RMSE	R^2
Linear Regression (Baseline)	4,167,285	4,899,759	-0.158
Ridge	3,381,503	4,032,661	0.216
Lasso	3,492,841	4,229,941	0.137

Final model performance is evaluated on a held-out test set using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics jointly assess average prediction accuracy, sensitivity to large errors, and overall explanatory power.

As reported in Table I, Ridge regression achieves the lowest prediction error and the highest R^2 among the evaluated models. This confirms that L2-regularized shrinkage provides a more stable fit when modeling log-salaries with high-dimensional TF-IDF features. Lasso regression exhibits slightly weaker performance, reflecting the trade-off between sparsity and predictive accuracy.

The baseline linear model performs poorly, with extremely large errors and a highly negative R^2, indicating numerical instability and failure to generalize. This result underscores the necessity of regularization when applying linear regression to text-based representations.

6. Prediction Fit: Actual vs Predicted Series

This section compares actual salaries (X-axis) and model-predicted salaries (Y-axis). The 45° dotted line indicates perfect calibration; points directly on the line mean that the prediction is equal to the actual value, points below the line indicate that the model underpredicts, and points above the line indicate overprediction. In general, the distribution of points shows that Ridge is closer to the diagonal line than Lasso, especially in the low to medium salary range. In the high salary range, both models tend to be more scattered and often fall below the line, indicating a tendency to underpredict for positions with premium compensation.

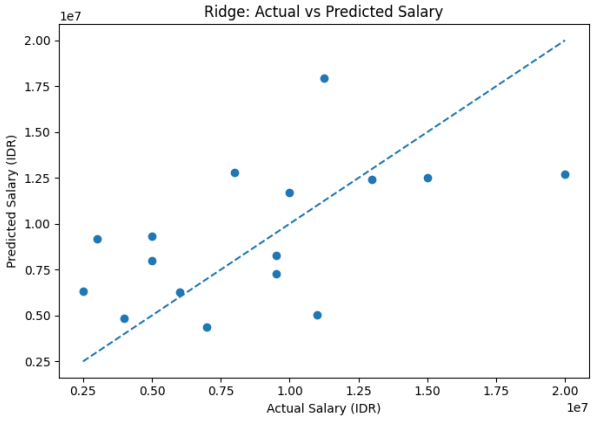


FIGURE 11 RIDGE: ACTUAL VS PREDICTED SALARY

The relatively compact distribution of points around the diagonal line in the majority of observations indicates better

calibration and smaller errors in the general salary range. The largest deviation appears at the right end (high salaries), where several points lie well below the line, indicating that the model is conservative towards high-paying vacancies or sensitive to outliers. Overall, this pattern is consistent with the more stable test performance on Ridge.

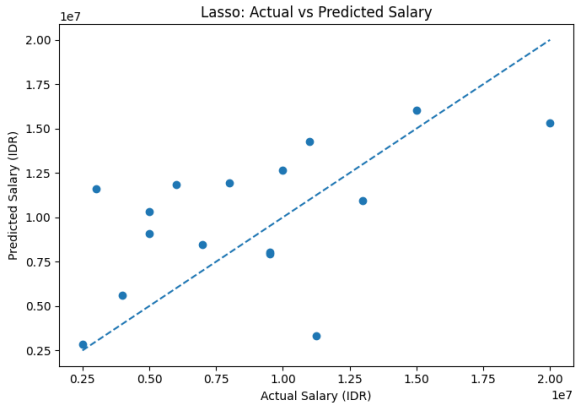


FIGURE 12 LASSO: ACTUAL VS PREDICTED SALARY

The distribution of points appears to be wider across the entire salary range, indicating greater error variance. At high salary levels, deviations from the diagonal line occur more frequently and are predominantly underpredictions. This pattern is consistent with the nature of Lasso, which reduces many coefficients to zero, so that signals spread across many tokens/phrases (e.g., a combination of technical skills and seniority) can be lost and accuracy decreases, especially for uncommon job opening

7. Error Diagnostics: Residual Analysis

We inspect residuals (actual − predicted salary) to assess model adequacy. For a well-specified linear model, residuals should be centered near zero with no obvious trend across the prediction range and with roughly constant spread. Systematic structure, skew, or variance that grows with the fitted value indicates specification issues (e.g., missing covariates, nonlinearity, heteroscedasticity) that can degrade inference and out-of-sample accuracy in the Indonesian postings context.

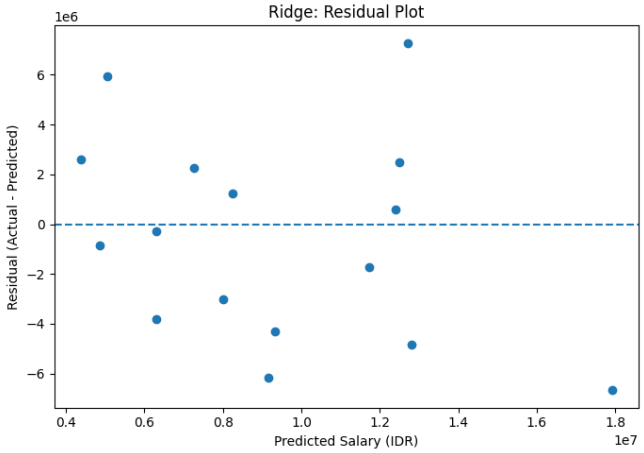


FIGURE 13 RIDGE: RESIDUAL PLOT

Residual analysis is conducted to assess the adequacy of the regression models and to detect potential violations of linear regression assumptions. Residuals are defined as the difference between actual and predicted salaries on the test set. For a well-specified linear model, residuals should be centered around zero, display no systematic trend with respect to fitted values, and exhibit approximately constant variance across the prediction range.

Figure X presents the residual plot for the Ridge regression model. The residuals are generally centered around zero, with both positive and negative errors distributed across the predicted salary range. No strong monotonic pattern or curvature is observed, suggesting that the linear specification is broadly appropriate for the log-transformed salary outcome. However, some dispersion increases at higher predicted salary levels, indicating mild heteroscedasticity. This pattern is common in wage data and likely reflects the limited number of high-paying job postings and unobserved heterogeneity in senior or specialized roles.

Figure Y shows the residual plot for the Lasso regression model. Compared to Ridge, the residuals appear more dispersed, particularly at higher predicted salaries. Several large positive and negative residuals indicate greater prediction errors for high-end postings. This behavior is consistent with Lasso's sparse feature selection, which may omit correlated textual signals that are relevant for capturing salary variation at the upper tail of the distribution.

Overall, the residual diagnostics indicate that Ridge regression produces more stable and homoscedastic error behavior than Lasso. While neither model fully eliminates heteroscedasticity, an inherent challenge in salary modeling the absence of strong systematic patterns supports the validity of the modeling approach. These findings further justify the selection of Ridge regression as the primary specification for analyzing salary determinants in Indonesian job postings.

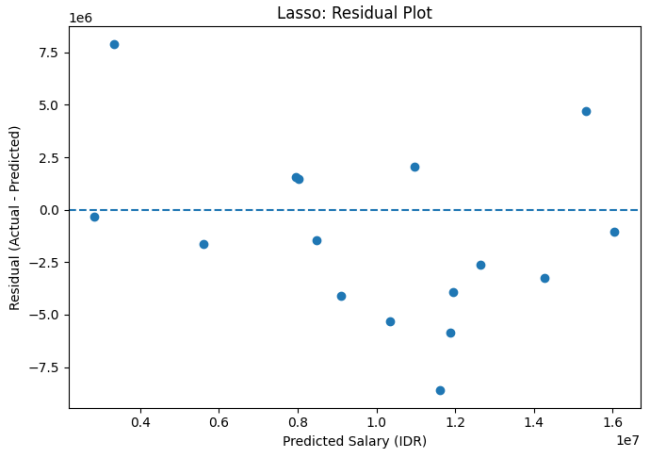


FIGURE 14 LASSO: RESIDUAL PLOT

The Lasso residual plot is a significant diagnostic tool for evaluating the generalisation capability of the sparse TF-IDF–based model in the context of From Words to Wages: The present study aims to explore the role of job description text mining in understanding salary determinants in Indonesia. In an ideally specified regression, residuals should be randomly scattered around the zero line with no systematic pattern, indicating that the model captures the main structure in the salary signal. In this figure, the residuals are not distributed uniformly: several points fall substantially below zero at mid-to-high predicted salary levels, while a few large positive residuals appear at the lower predicted range. This dispersion suggests that the Lasso model is unstable in its capacity to capture salary variation across the full compensation spectrum, particularly in the context of predicting postings associated with higher pay.

A salient feature is the presence of substantial negative residuals surrounding the mid-to-upper predicted range (approximately between 1.0×10^7 and 1.5×10^7 IDR), denoting instances where the model overestimates salaries relative to the actual disclosed values. Conversely, at the lower end of the predicted salary range, at least one observation displays a substantial positive residual, indicating that the model underestimates the true salary for that specific posting. This asymmetric behaviour is consistent with a sparse model that relies on a limited subset of tokens. When a job advertisement contains a few high-weight phrases that Lasso retains, the prediction can become overly sensitive to those terms, producing overprediction or underprediction depending on whether the retained vocabulary accurately represents the economic value of that job. In practical terms, this implies that certain linguistic signals in Indonesian job descriptions may be highly context-dependent, and Lasso may fail to capture nuance when relevant terms are correlated or when salary-relevant meaning is expressed through varied phrasing.

From an interpretability perspective, these residual patterns also align with the known trade-off of Lasso in high-dimensional text settings. Lasso's methodology for performing feature selection involves the shrinkage of numerous coefficients to zero, thus enabling the algorithm to disregard correlated wage signals that manifest as alternative phrasings (e.g., divergent ways of articulating seniority, leadership responsibilities, or tool requirements). In bilingual Indonesian postings, where similar competencies may be expressed in English or Bahasa Indonesia, sparsity can further amplify representation gaps and weaken robustness. Therefore, while Lasso is attractive for producing a concise set of "determinant words," the residual diagnostics suggest that in this dataset it may not provide the most stable salary estimation framework, especially at the upper tail of the wage distribution.

The Lasso residual plot indicates that salary determination signals extracted from job-description text are present, but that these signals are learnt unevenly by a sparse model. The objective of the present study is twofold: firstly, to emphasise both predictive performance and credible interpretation of salary determinants; and secondly, to utilise Ridge regression as the primary model for inference, while positioning Lasso as a complementary analysis for feature screening. The residual behaviour supports the use of Ridge regression, while the results of the study support the use of Lasso for feature screening. This strategy preserves interpretability without sacrificing stability, allowing the paper to more reliably translate "words" in Indonesian job advertisements into measurable "wage" determinants across regions and seniority levels.

8. Interpretation of Salary Determinants

We interpret model coefficients as marginal effects on the target (monthly salary, in IDR) holding other features constant. Positive coefficients indicate terms/attributes associated with higher offered pay; negative coefficients indicate the opposite. Because Ridge distributes weight across correlated features while Lasso enforces sparsity, we read the two plots together: Ridge highlights broad, stable signals; Lasso points to a smaller set of dominant predictors (and potential red flags) that survive strong shrinkage.

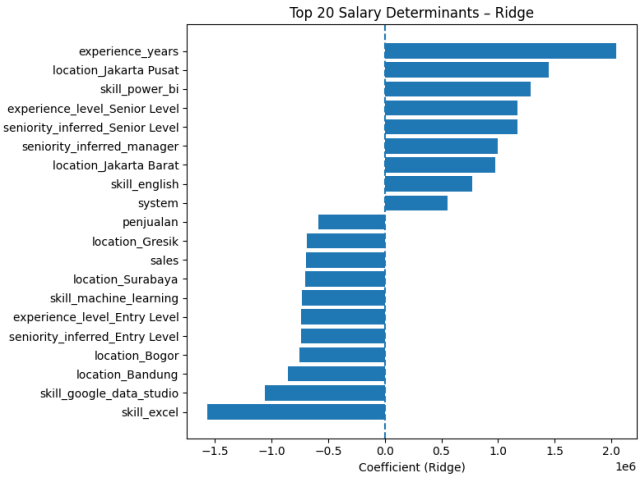


FIGURE 15 TOP 20 SALARY DETERMINANTS – RIDGE

The Ridge coefficient plot reveals a coherent labor market pattern consistent with established wage-setting mechanisms. Years of experience emerge as the strongest positive predictor of salary, aligning with standard returns-to-experience documented in Indonesian data and analytics occupations. Location effects are also substantial: indicators for Central Jakarta and West Jakarta exhibit large positive coefficients, reflecting a clear wage premium in Jakarta's core labor markets relative to other cities.

Seniority-related variables—including Senior level indicators and inferred managerial roles—are strongly associated with higher predicted salaries. This pattern suggests that language signaling greater responsibility, leadership, or decision-making authority is systematically rewarded in compensation. Among skill-related features, indicators for analytical proficiency and English language usage carry positive coefficients, reflecting higher demand for client-facing and internationally oriented roles.

Conversely, variables proxying entry-level or junior profiles display negative associations with salary. Indicators for entry-level experience and inferred junior seniority consistently reduce predicted compensation, consistent with lower wage offers for early-career positions. Several location dummies outside Jakarta, such as Bandung and Bogor, and to a lesser extent Surabaya and Gresik, exhibit negative coefficients, aligning with lower median salary levels in non-core urban areas.

At the token level, commonly listed tools such as spreadsheet software and basic analytics platforms show negative associations with salary, likely capturing baseline or entry-level tooling requirements. In contrast, the modest negative coefficient on machine learning-related terms may reflect multicollinearity with seniority and location features, as well as limited sample representation. Because Ridge regression distributes weight across correlated predictors, individual coefficients should be interpreted jointly rather than in isolation.

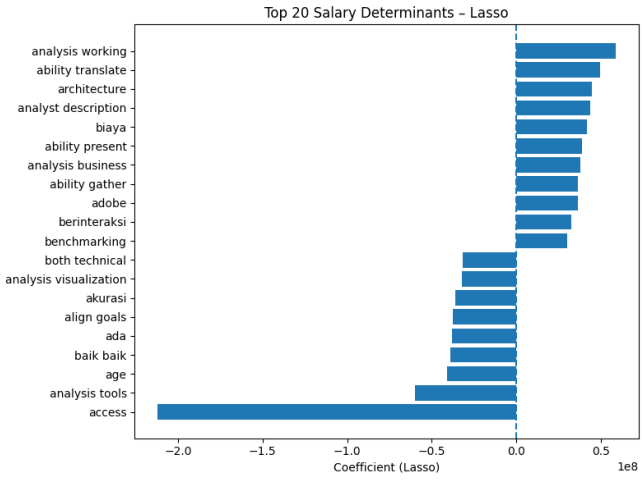


FIGURE 16 TOP 20 SALARY DETERMINANTS – LASSO

Lasso highlights a sparse set of terms with large magnitudes. On the positive side, phrases such as 'analysis working', 'ability translate', 'architecture', 'analyst description', and 'analysis business' point to postings that emphasise systems architecture, translation of business needs, and presentation ability, signals of hybrid technical roles that tend to pay more.

Conversely, Lasso flags a few large-magnitude negatives, most notably the word 'access' and phrases such as 'analysis tools' and 'age'. These are likely to be artefacts of phrasing or segments of the market where the same words co-occur with junior tasks, legacy stacks or compliance language, which drives down predicted pay. As Lasso eliminates many correlated features, the surviving tokens can absorb the effect of entire clusters. Therefore, extremely large positive or negative weights should be validated against frequency, context windows and bilingual variants, to rule out spurious correlation

When considered as a whole, both models concur that experience, seniority signals, central-Jakarta locations, English language proficiency, and business-facing analytics tools are associated with higher offers, while entry-level flags, non-Jakarta locations, and baseline/spreadsheet-heavy skill cues correspond to lower offers in this dataset. Ridge provides a stable, interpretable ranking; Lasso surfaces a compact list of salient phrases but can exaggerate idiosyncratic tokens. It is recommended that (i) semantically similar skills be grouped, (ii) coefficient stability be checked across folds/time, and (iii) token contexts be inspected to ensure that linguistic artefacts (e.g. bilingual stemming, boilerplate) are not misread as economic signs.

V. CONCLUSION

This study investigates how unstructured job-description text and structured job attributes jointly shape salary outcomes in the Indonesian labor market. By applying text mining techniques to online job postings and integrating TF-IDF representations with regularized regression models, the paper demonstrates that wage-relevant information is systematically embedded in the language of job advertisements.

The empirical results show that disclosed salaries exhibit substantial right-skewness and heterogeneity across seniority levels and locations, motivating the use of log-transformed salaries and regularization for stable estimation. Exploratory analysis confirms clear compensation gradients by inferred seniority, while cross-validation results indicate that Ridge regression provides more robust and consistent predictive performance than both the baseline and Lasso models in a high-dimensional text setting.

Beyond prediction accuracy, coefficient-based interpretation reveals economically meaningful salary determinants. Experience, seniority signals, and Jakarta-based locations consistently carry positive associations with compensation, reflecting standard labor market mechanisms such as returns to experience and urban wage premiums. Skill-related terms and language indicators also contribute to salary variation, confirming that job-description wording conveys information about role complexity and market value. Conversely, entry-level signals and baseline tooling requirements are associated with lower predicted salaries, consistent with junior labor market positioning.

Overall, the findings support the central premise that “words matter for wages.” Job postings do not merely describe roles but encode structured signals that correlate strongly with compensation. By combining interpretable text features with regularized regression, this study contributes to both labor economics and applied data science in the

Indonesian context. Future research may extend this framework by incorporating richer semantic embeddings, panel data across time, or causal identification strategies to further disentangle wage-setting mechanisms in online labor markets.

REFERENCES

[1] **Badan Pusat Statistik, "Upah Rata-Rata per Jam Pekerja Menurut Jenis Kelamin, 2023," 2024.**

[2] **World Bank, "Gender Equality for Growth: Research and Analytical Program in Indonesia," Country Brief, 2025.**

[3] **A. Pisolla, S. B. Moghul, T. Gaddam, dan N. C. S. N. Iyengar, "Salary prediction using TF-IDF and ensemble machine learning: A lightweight and interpretable approach," *World Journal of Advanced Research and Reviews*, vol. 26, no. 2, pp. 4445–4453, May 2025.**

[4] **L. Chen et al., "An extended TF-IDF method for improving keyword extraction," *Data & Knowledge Engineering*, vol. 152, pp. 1–10, 2022.**

[5] **"Data science for job market analysis: A survey on applications and techniques," *Expert Systems with Applications*, 2021, doi: 10.1016/j.eswa.2021.115544.**

[6] **A. R. Alsheyab, M. Alkasawneh, and N. Shahin, "Job Market Cheat Codes: Prototyping Salary Prediction and Job Grouping with Synthetic Job Listings," 2025. [Online]. Available: link in table.**

[7] **H. R. Qaiser and R. K. Rajpoot, "Text mining: use of TF-IDF to examine the relevance of keywords to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.**

[8] **T. Lesmana, "Impact of artificial intelligence on job transfer and workforce adaptation in start-up companies in Indonesia," *Eastasouth Proceedings of Nature, Science, and Technology*, vol. 1, no. 01, pp. 63–73, June 2024.**

[9] **H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.**

[10] **Z. Ao, G. Horváth, C. Sheng, Y. Song, and Y. Sun, "Skill requirements in job advertisements: A comparison of methods," *Information Processing & Management*, 2023.**

[11] **J. Vogt, T. Voigt, A. Nowak, and J. M. Pawlowski, "Development of a Job Advertisement Analysis for Assessing Data Science Competencies," 2023.**

[12] **M. den Hartog, S. Bras, and G.-J. Schoenmakers, "The Impact of Pronouns of Address in Job Ads from Different Industries and Companies," 2024.**

[13] **R. Audoly, M. Bhuller, and T. A. Reiremo, "The Pay and Non-Pay Content of Job Ads," 2024**

[14] **S. Fareri, G. Fantoni, F. Chiarello, E. Coli, and A. Binda, "Estimating Industry 4.0 impact on job profiles and skills using text mining," 2020.**

[15] **R. Wang, Q. Chen, Y. Wang, B. Shen, and L. Xiong, "JobViz: Skill-driven visual exploration of job advertisements," *Visual Informatics*, 2024.**

[16] **I. Rahhal, I. Kassou, and M. Ghogho, "Data science for job market analysis: A survey on applications and techniques," *Expert Systems with Applications*, vol. 251, Art. 124101, 2024. DOI: 10.1016/j.eswa.2024.124101.**

[17] **G. Tzimas, N. Zotos, E. Mourelatos, K. C. Giotopoulos, and P. Zervas, "From Data to Insight: Transforming Online Job Postings into Labor-Market Intelligence," *Information*, vol. 15, no. 8, Art. 496, 2024. doi: 10.3390/info15080496.**

[18] **Y. Sun, F. Zhuang, H. Zhu, Q. Zhang, Q. He, and H. Xiong, "Market-oriented job skill valuation with cooperative composition neural network," *Nature Communications*, vol. 12, Art. number 1992, 2021. DOI: 10.1038/s41467-021-22215-y.**

[19] **G. Airlangga, "Performance evaluation of machine learning models for predicting household energy consumption: A comparative study," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 8, no. 1, pp. 76–85, Mar. 2025, doi: 10.24014/ijaidm.v8i1.32791.**