PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol 04 No 1, April 2024
Page 11-28

IDENTIFICATION OF DOMINANT FACTORS IN STUNTING CASE NARRATIVES USING PCA AND SVD APPROACHES

Debora Chrisinta[1]*, Justin Eduardo Simarmata[2], Milagros R. Baldemor[3], Miko Purnomo[4]

[1]Information Technology Study Program, Faculty of Agriculture, Science and Health, University of Timor
El Tari Street - Km. 09 Sasi , Kefamenanu City, TTU, 85613, NTT, Indonesia
[2]Mathematics Education Study Program, Faculty of Teacher Training and Education, University of Timor
El Tari Street - Km. 09 Sasi , Kefamenanu City, TTU, 85613, NTT, Indonesia
[3]Mathematics and Allied Disciplines Department, Don Mariano Marcos Memorial State University
National Highway, Barangay Catbangen, City of San Fernando, 2500, La Union, Philippines
[4]Mathematics Study Program, Faculty of Agriculture, Science and Health, University of Timor
El Tari Street - Km. 09 Sasi , Kefamenanu City, TTU, 85613, NTT, Indonesia

E-mail Correspondence Author: *deborachrisinta@unimor.ac.id*

Abstract

Stunting remains a serious public health concern in Indonesia, exacerbated by the limited public understanding of its causes and prevention strategies. This study analyzes public perceptions of stunting based on reviews collected through web scraping from the 2023 Indonesian Health Survey (SKI). Text preprocessing techniques, Term Frequency-Inverse Document Frequency (TF-IDF) analysis, and dimensionality reduction using Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) were applied to a dataset comprising 21 reviews. The results indicate that PCA outperforms SVD in simplifying the relationships among key terms, as evidenced by a lower reconstruction error (0.003861 compared to 0.004232). The dominant factors influencing public perception include education, sanitation, and socio-economic conditions. These findings highlight the critical role of data-driven and visual-based educational strategies in enhancing public awareness and accelerating stunting prevention efforts.

Keywords: *Public Perception, PCA, Stunting, SVD, Text Analysis, TF-IDF*

: https://doi.org/10.30598/parameterv4i1pp11-28

1. INTRODUCTION

In the current digital era, various opinions and perspectives on health issues, including stunting, are widely published on numerous website platforms. The narratives formed from public opinion can serve as valuable sources of information for understanding the behavior of communities affected by stunting. Analyzing these narratives can aid in formulating more targeted policies for stunting mitigation efforts. Given the vast availability of online narratives, effective methods are required to process and extract relevant information from the available text **[1]**. Advancements in natural language processing (NLP) and big data analytics have enabled deeper exploration of hidden information within textual data. These technologies allow researchers to identify patterns, trends, and dominant factors in text more efficiently than traditional approaches. Consequently, the application of machine learning-based data analysis techniques and multivariate statistics is highly relevant for processing large volumes of unstructured text data **[2]**.

This study focuses on analyzing narratives found on the website https://dashboard.stunting.go.id/persepsi-masyarakat-tentang-stunting/ to identify dominant factors emerging in discussions about stunting cases. To collect data, a web scraping approach is employed to extract information from the website. This technique enables the automated and systematic retrieval of data from various online sources. The scraping method used in this study follows the approach outlined in previous research **[3]**, which explores web scraping techniques in Python-based data mining.

Previous studies have demonstrated the effectiveness of text analysis and web scraping across various domains. Chrisinta and Simarmata conducted a comparative study between Support Vector Machine and Naïve Bayes for sentiment analysis on lecturer performance, revealing that machine learning techniques can uncover hidden patterns within textual data **[4]**. Text mining analysis on agroforestry using web scraping and topic modeling, proving that this approach can extract valuable insights from unstructured data **[5]**. Additionally, **[6]** applied a collaborative filtering approach based on Singular Value Decomposition (SVD) for skincare product recommendations, highlighting the superiority of SVD in handling text-based recommendation data. Furthermore, **[7]** emphasized the potential of web mining in predictive business analytics, demonstrating that this technique can transform raw data into actionable insights.

Identifying dominant factors in narratives related to stunting cases is crucial, as it provides deeper insights for policymakers and relevant organizations in formulating more effective communication and intervention strategies. Previous studies have highlighted the benefits of factor analysis in understanding public perceptions of health issues **[8]** **[9]**. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are widely used methods for dimensionality reduction in textual data, enabling the extraction of key information without losing the essence of the original data. These methods were selected for this study because they effectively identify primary patterns in large and complex textual datasets. PCA reduces data dimensionality while preserving significant variability, whereas SVD decomposes text into a simpler yet meaningful representation **[10]**[**11]**. However, each method has its advantages and limitations. PCA excels in simplifying data while retaining essential information but may face challenges when handling extremely large and structurally complex textual data. On the other hand, SVD is highly effective for processing large matrices and uncovering latent relationships in textual data but can be computationally demanding and prone to overfitting if not properly applied. Therefore, further research is needed to evaluate the performance of

these methods in analyzing narratives on stunting cases and to determine the most suitable approach for this context.

Several previous studies have employed these methods in text analysis, such as document classification and topic clustering. However, this study differs by focusing on the analysis of stunting case narratives and the social context in Indonesia. The objective of this research is to identify the dominant factors in public narratives regarding stunting, as presented on the specified website. The findings of this study are expected to provide broader insights into the open narratives available on the website concerning stunting cases. Additionally, the results may serve as a foundation for developing more effective policies in addressing this issue.

Despite the growing application of text mining techniques in various fields, limited research has specifically explored the use of PCA and SVD in analyzing public narratives related to stunting, particularly in the Indonesian context. Existing studies often focus on sentiment classification or topic modeling in more general domains, leaving a gap in leveraging dimensionality reduction techniques to uncover key factors influencing public perceptions of stunting. Therefore, this study aims to address this gap by applying PCA and SVD to systematically identify dominant factors from online narratives about stunting. The specific objectives are (1) to evaluate the performance of PCA and SVD in extracting critical factors from stunting-related text data and (2) to provide actionable insights that can inform public health strategies and policy-making processes targeting stunting prevention in Indonesia.

2. RESEARCH METHODOLOGY

2.1. Research Design

This study employs a quantitative approach with an exploratory method aimed at identifying dominant factors in stunting case narratives through text analysis. A quantitative approach was chosen because this research involves data processing using multivariate statistical techniques and machine learning to extract patterns from text objectively **[2][12]**. The exploratory method is utilized as this study seeks to uncover hidden factors within textual data without a strict initial hypothesis. Text data exploration techniques, such as those applied by **[5]** in agroforestry topic analysis, have proven effective in revealing unexpected insights from unstructured data. Additionally, this approach was implemented by **[13]** in sentiment analysis of lecturer performance, demonstrating the effectiveness of quantitative methods in machine learning-based text analysis. Given the relatively small sample size (21 reviews), the findings are considered preliminary and exploratory, warranting cautious interpretation.

2.2. Sources and Techniques of Data Collection

2.2.1 Data Sources

This study utilizes secondary data obtained from the official website https://dashboard.stunting.go.id/persepsi-masyarakat-tentang-stunting/. This data source was selected because it provides various narratives from the public regarding perceptions of stunting cases in Indonesia. The data includes opinions, experiences, and public perspectives, which are valuable for analysis to identify dominant factors frequently appearing in discussions on stunting. The use of secondary data in text analysis has been a widely adopted approach in previous research. According to **[14]**, secondary data from various digital platforms can offer extensive insights for analyzing social and

business trends. Furthermore, a study by [5] demonstrated that text exploration from online sources could provide in-depth information on public perception patterns across various fields, including health and social issues.

2.2.2 Data Collection Techniques

Data is collected using the web scraping method, an automated technique for extracting information from web pages. Previous research has demonstrated the effectiveness of web scraping for text analysis. [15] revealed that Python-based web scraping techniques can accurately and optimally extract textual data in the context of data mining.

2.3 Data Processing Methods

The data processing method aims to ensure that the data used in this study is in optimal condition for analysis. This stage includes data cleaning, transforming text into a computationally processable format, and representing data as numerical vectors. This process is crucial for improving the accuracy of models in extracting information from textual data [16].

2.3.1 Data Preprocessing

Data preprocessing is an initial step in text processing aimed at cleaning and simplifying text for more efficient analysis. Several key techniques in data preprocessing include:

a) Stopwords removal, which eliminates common words that do not carry significant meaning in text analysis, such as "and," "that," or "or" [1].
b) Stemming, which converts words to their root forms by removing suffixes, for example, "berlari" becomes "lari" [17].
c) Lemmatization, which converts words to their base forms while considering linguistic context, for example, "running" becomes "run" [18].
d) Tokenization, which breaks text into word or phrase units for further analysis [17].

2.3.2 Data Representation

After preprocessing, text must be represented in a numerical format for further analysis. One commonly used method is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a technique for measuring the importance of a word in a document relative to the entire corpus. The formula is as follows:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

where $TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$ is measures how often a term t appears in a document d, $IDF = log\frac{N}{1+n_t}$ is evaluates how important a term is within the entire document collection, $f_{t,d}$ is he number of times term t appears in document d, $\sum_{t' \in d} f_{t',d}$ is the total number of terms in d, N is the total number of documents, n_t is the number of documents containing term t [19].

2.4 Data Processing Methods

The data analysis method aims to extract essential information from textual data represented using TF-IDF. In this study, two primary methods are employed to identify dominant factors in stunting case narratives: Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). These methods are used for dimensionality

reduction and to uncover hidden patterns in textual data [20][21]. Additionally, Graph Visualization is utilized to illustrate the relationships among the dominant factors identified, providing a clearer understanding of the structure and connections within the narrative data.

2.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used to identify the principal components in a dataset while preserving as much information as possible. This method operates by transforming the data into a lower-dimensional space using a linear combination of the original variables [22]. The fundamental equation of PCA is as follows:

$$\boldsymbol{Z} = \boldsymbol{XW} \tag{2}$$

where $\boldsymbol{X}$ is a data matrix of size $m \times n$, where m represents the number of samples and n represents the number of features. $\boldsymbol{W}$ is the eigenvector matrix (principal components), and $\boldsymbol{Z}$ is the transformed matrix in the new dimensional space. PCA is widely used in text analysis to identify hidden patterns and reduce data dimensionality without losing essential information [11].

2.4.2 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a matrix decomposition technique used to uncover hidden structures in data by decomposing a matrix into three main components:

$$\boldsymbol{A} = \boldsymbol{U\Sigma V^T} \tag{3}$$

where $\boldsymbol{A}$ is a data matrix of size $m \times n$, $\boldsymbol{U}$ is an orthogonal matrix representing features in a lower-dimensional space, $\boldsymbol{\Sigma}$ is a diagonal matrix containing singular values, and $\boldsymbol{V^T}$ is an orthogonal matrix representing relationships between variables. SVD is widely used in text mining and factor analysis due to its ability to capture hidden patterns in textual data [12].

2.4.3 Graph Visualization

Graph visualization is used to illustrate the relationships between the dominant factors obtained from PCA and SVD. This technique enables a more intuitive analysis by presenting the connections between elements in a graphical format. In this study, a network graph approach is employed to depict the proximity of dominant factors based on their co-occurrence in documents. The visualization model follows a force-directed graph approach, where factors with stronger relationships are positioned closer to each other in the graphical representation [23]. The fundamental equation used in graph visualization is as follows:

$$F = k\left(\frac{d_{ij}^2}{r}\right) \tag{4}$$

where F represents the attractive force between nodes, k is a constant that depends on the scale and system used, $d_{ij} = \sqrt{\sum(x_i - x_j)^2}$ is the distance between factors in the data, also known as the Euclidean distance, dan r represents the weight or relevance factor between two entities (e.g., dominant factors in text analysis).

2.5 Model Evaluation

Model evaluation aims to assess the effectiveness of Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) in identifying dominant factors in stunting case narratives. The evaluation process includes measuring model performance and validating the results by comparing the identified dominant factors with data from the original source. A sensitivity analysis was conducted by varying the number of retained components/singular values to test the robustness of the findings, ensuring consistency despite small data fluctuations.

2.5.1 Performance Measurement

Measuring the explained variance for different numbers of principal components (in PCA) or singular values (in SVD) and plotting the results. The elbow point is identified where the variance gain starts to plateau, indicating that adding more components or singular values no longer significantly improves information retention. In PCA and SVD, each principal component or singular value explains a portion of the variance in the data. Selecting too few components may result in the loss of critical information, whereas selecting too many can lead to an unnecessarily complex model with minimal added benefits. The elbow method approach is implemented by:

a) Calculating the eigenvalues from the covariance matrix in PCA or singular values in SVD.
b) Sorting the eigenvalues or singular values in descending order.
c) Computing the variance explained ratio for each component or singular value using the formula:

$$Variance\ Explained = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{5}$$

where λ_i represents the i-th eigenvalue (in PCA) or singular value squared (in SVD), and $\sum_{j=1}^{n} \lambda_j$ is the sum of all eigenvalues or squared singular values.

d) Plotting the number of components vs. variance explained to visualize how much information is retained as more components or singular values are added.
e) Identifying the elbow point, which is the point where adding more components no longer provides a significant increase in the explained variance.

Reconstruction error is used to evaluate the quality of dimensionality reduction in PCA and SVD **[24]**. This method measures the extent of information loss after the data is reduced to a certain number of components or singular values. In PCA, the reconstruction error indicates the extent to which the original data can be reconstructed from the selected principal components. Meanwhile, in SVD, the reconstruction error measures how well the original matrix can be reconstructed from the chosen number of singular values. The equation used is as follows:

$$RE = ||A - A_k||_F \tag{6}$$

where A represents the original data matrix, $A_k = U_k \Sigma_k V_k^T$ is the reconstructed matrix using k components or singular values, and $||\cdot||_F$ denotes the Frobenius norm, which is defined as:

$$\left|\left|X\right|\right|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} \tag{7}$$

The optimal reconstruction error value is selected when the addition of more components no longer results in a significant reduction in error.

2.5.2 Performance Measurement

The validation of results aims to assess whether the dominant factors identified using PCA and SVD align with the narrative patterns present in the original dataset. The validation process is conducted by comparing the analysis results with the data obtained from the website.

2.6 Tools and Software Used

The primary programming language used in this research is Python, as it offers a wide range of libraries that support data mining, text analysis, and machine learning. The libraries utilized in this study are presented in **Table 1** below.

Table 1. Python Libraries Used

Category	Library	Function
Data Collection	BeautifulSoup, Scrapy, Requests	Retrieving text data from websites through web scraping techniques.
Text Processing	NLTK, spaCy, TextBlob	Tokenize, stemming, lemmatization, and stopword removal.
Data Representation	Scikit-learn, TF-IDFVectorizer	Convert text to numerical representations with TF-IDF.
Data Analysis	Scikit-learn, NumPy, SciPy	Apply PCA and SVD for feature extraction and dimension reduction.
Visualization	Matplotlib, Seaborn, Plotly, NetworkX	Displays the results of the analysis in the form of graphs, including graph visualization.

2.7 Research Flow

The following outlines the research stages in identifying dominant factors referenced in stunting cases based on narrative data from a website discussing stunting:

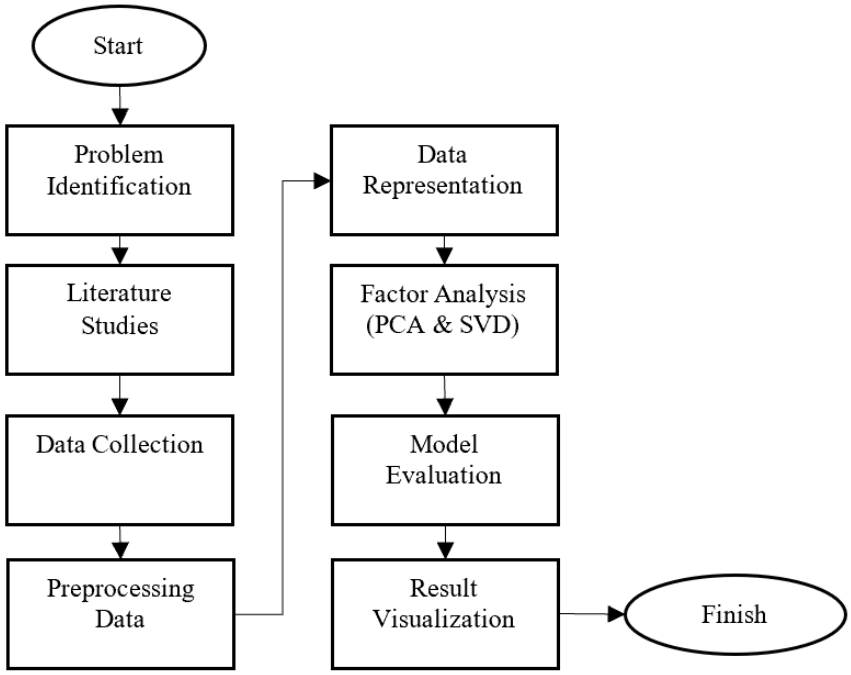

Figure 1. Research Flow Diagram

3. RESULTS AND DISCUSSION

3.1 Scraping and Data Preprocessing Results

The initial results from the text scraping process related to public perceptions of stunting are presented in **Table 2**. The extracted data consist of sentences taken from reviews on the website, reflecting public understanding of stunting based on the 2023 Indonesian Health Survey (SKI). A total of 21 sentences were obtained from the website reviews.

Table 2. Scraped Data

Index	Sentences
0	Based on the SKI 2023 data, 69.40% of the community has a correct understanding of stunting, while 30.60% still have a wrong understanding.
1	This shows that the majority of people have understood important aspects related to stunting, such as the causes, impacts and ways to prevent it.
2	However, the percentage who still have a wrong understanding is quite significant, almost a third of the total respondents.
3	The results are shown in the pie chart below.
⋮	⋮
21	With more intensive health campaigns, it is hoped that the stunting rate can continue to be reduced.

After the text preprocessing stage, the results are presented in Table 3 for conversion into numerical data, enabling the application of PCA and SVD data analysis methods.

Table 3. Preprocessed Data

Index	Sentences
0	based on ski data as much as the community has an understanding of stunting while still having a wrong understanding
1	shows that the majority of people have understood important aspects related to stunting such as the causes of the impact of prevention
2	but the percentage that still has a significant misunderstanding is one-third of the total respondents
3	the results are as shown in the pie chart next to
⋮	⋮
21	more intensive health campaigns are expected to continue to reduce stunting rates

3.2 TF-IDF Analysis

After data preprocessing, the next step is to apply Term Frequency-Inverse Document Frequency (TF-IDF) to measure the importance of each word within the text corpus. The TF-IDF results are presented in Table 4, which consists of 22 rows (representing the number of reviews) and 134 columns (representing the number of unique words after preprocessing, used as features). In the table, each cell contains a TF-IDF value that reflects the significance of a word in a specific document (review) relative to the entire text corpus. A value of 0.000000 indicates that the word does not appear in the given review. Conversely, the higher the TF-IDF value, the more important the word is in the corresponding text.

Table 4. TF-IDF Results

Index	angka	aspek	bahwa	berdasarkan	...	total
0	0.000000	0.000000	0.000000	0.256349	...	0.000000

1	0.000000	0.324458	0.259125	0.000000	...	0.000000
2	0.000000	0.000000	0.000000	0.000000	...	0.352137
3	0.000000	0.000000	0.000000	0.000000	...	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0.354063	0.000000	0.000000	0.000000	...	0.000000

3.3 Implementation of PCA and SVD in Dimensionality Reduction

The contribution of components in the model is illustrated in Figure 2, which displays the contribution of each component in both PCA and SVD models. The graph on the left represents the proportion of variance explained by each principal component in PCA. It is evident that the first component has the highest contribution, indicating that the primary factor identified in the dataset has a dominant influence. Subsequent components exhibit a decreasing contribution to the explained variance. The graph on the right presents the singular values of the components in SVD. The highest singular value is observed in the first component, followed by a gradual decline in the subsequent components. This suggests that the primary dimensions of the data can be represented with fewer components without significant loss of information.

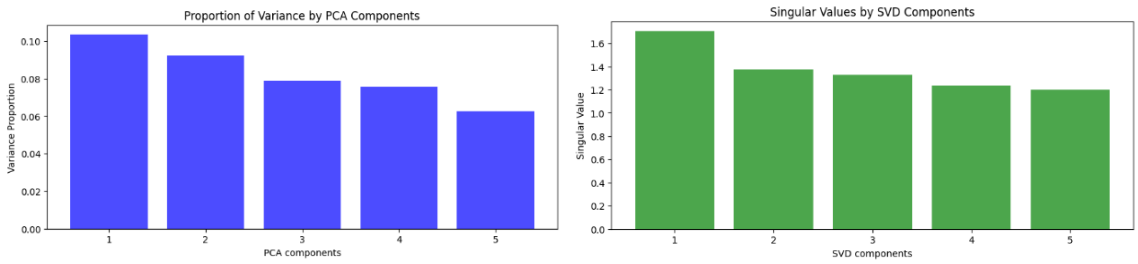

Figure 2. Component Contributions to PCA and SVD Models

Based on the word cloud visualization of the 20 highest-weighted words from the PCA and SVD models, key terms with the highest contribution in the analysis are identified (Figure 3). The PCA word cloud (left image) highlights words such as *memiliki*, *kembang*, *pemantauan*, *anak*, *tumbuh*, and *edukasi* as the primary contributing factors in the analysis. These words indicate that child growth monitoring and education play a crucial role in understanding stunting. Meanwhile, the SVD word cloud (right image) features words such as *pada*, *kesadaran*, *terlihat*, *chart*, *pemantauan*, and *edukasi*. This suggests that public awareness of stunting, along with the importance of education and monitoring, emerges as the dominant factors in the SVD-based analysis.

Figure 3. Component Contributions to PCA and SVD Models

3.4 Word Proximity Analysis Using Graph Visualization

In Figure 4 (PCA), several distinct word clusters are evident, such as *kemiskinan*, *masyarakat*, *pemahaman*, and *terkait*, which form the core of major relationships. These clusters indicate that these words frequently co-occur within the same context. Additionally, words like *stunting*, *tumbuh*, *kembang*, *anak*, and *gizi* create a group focused on health and child development issues. PCA appears to structure closely related words into more distinct clusters; however, the number of word connections is relatively limited compared to SVD. In Figure 5 (SVD), a greater number of connections between words is observed, suggesting that SVD captures a broader range of semantic relationships. For instance, the words *bar* and *pie* appear as part of a larger graph structure, possibly indicating their relevance to visual representation or data analysis in this study. Furthermore, words like *menyadari*, *mengetahui*, *pentingnya*, and *perlunya* are more dispersed and connected to multiple other words, highlighting SVD's ability to capture a wider meaning network compared to PCA.

Overall, PCA tends to produce more distinct and well-separated word clusters, while SVD results in a more complex and richly connected structure of word relationships. This suggests that SVD is more effective in preserving semantic relationships within textual data, whereas PCA is more efficient in simplifying word relationship structures.

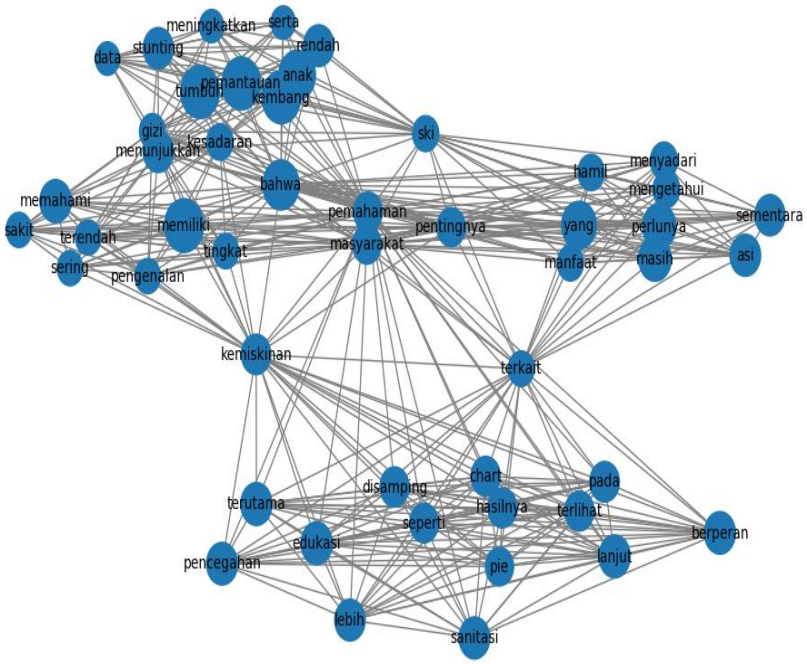

Figure 4. **Word Relationship Graph for PCA**

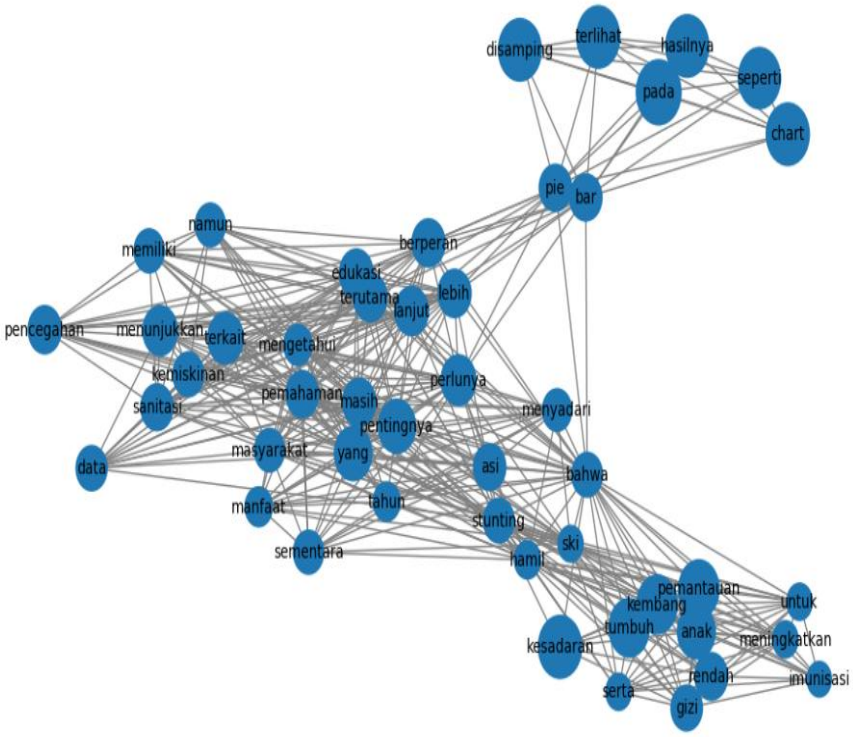

Figure 5. **Word Relationship Graph for SVD**

The graph visualization results using PCA and SVD indicate that key factors in discussions about stunting are related to public awareness, education, and socio-economic factors such as poverty and sanitation. PCA identifies clearer word clusters, highlighting the strong relationship between public understanding and socio-economic conditions,

while SVD captures broader semantic relationships, suggesting that public awareness of stunting is still evolving. These findings align with previous research by [25], which stated that poverty and lack of health education are major factors contributing to the prevalence of stunting in various developing countries. Additionally, research by [26] also emphasized that better knowledge of nutrition and child growth through educational campaigns can significantly reduce stunting rates. The differences in relationship patterns observed in PCA and SVD suggest that effective education and visual data representation can play a crucial role in increasing public awareness of stunting prevention.

3.5 Evaluation of PCA and SVD Models

In Figure 6 presents the model evaluation of PCA and SVD based on the number of selected components. In the left graph, the PCA evaluation is shown as Cumulative Variance Explained, indicating that the more components selected, the greater the variance explained by the model. However, at a single-component point (marked by the red dashed line), PCA can only explain approximately 10% of the variance in the data, with a reconstruction error of 0.003861. Meanwhile, in the right graph, the SVD evaluation is displayed using Singular Values, illustrating the contribution of each component to the data structure. The singular values in SVD gradually decrease, with the first component having the highest singular value of around 1.7 and a reconstruction error of 0.004232. These results indicate that while SVD captures more semantic relationships within the data (as observed in Figure 5), PCA is more efficient in representing information with fewer components. A previous study by [12] on Latent Semantic Analysis (LSA) demonstrated that SVD is superior in capturing latent semantic relationships in text compared to PCA-based methods. Additionally, [27] highlighted that PCA is highly effective in dimensionality reduction with minimal information loss but is less flexible than SVD in capturing indirect word associations. Thus, the findings of this study confirm previous research, suggesting that PCA is more suitable for simpler dimensionality reduction tasks, while SVD is more robust in preserving semantic relationships within textual data.

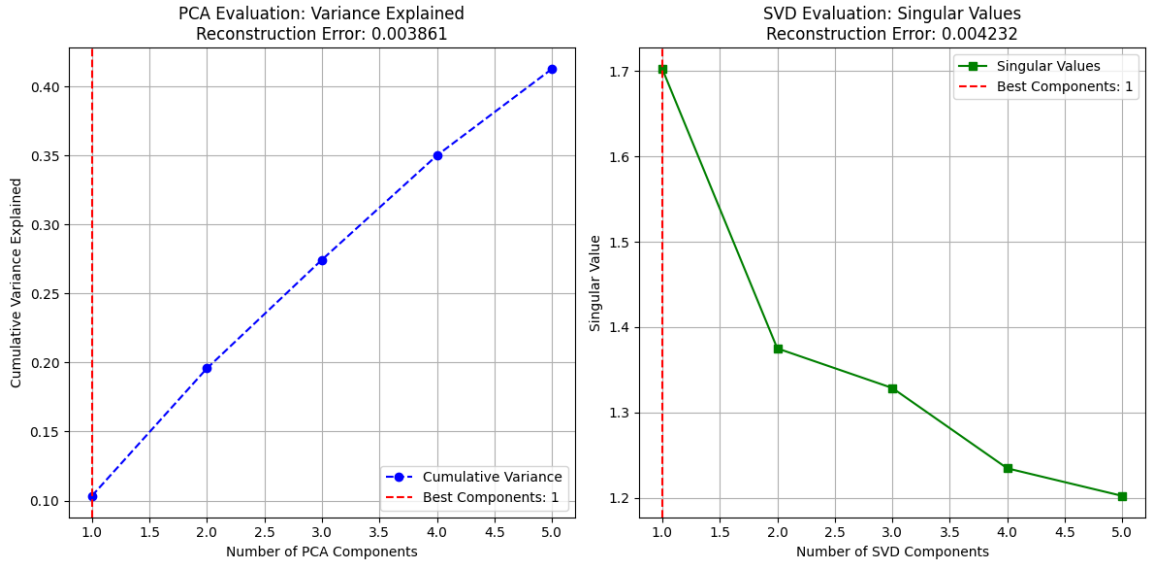

Figure 6. Goodness-of-Fit Plot of PCA and SVD Models

3.6 Implications of Findings and Recommendations

In **Figure 7** presents the Graph of Key Factors based on dimensionality reduction using PCA, which identifies relationships between key terms in the text. Words with

larger node sizes, such as *seperti* (like), *edukasi* (education), *lebih* (more), and *hasilnya* (result), indicate the primary factors that frequently appear in discussions regarding public understanding of stunting. The connections between words are represented by lines of varying weights, which illustrate semantic proximity within the dataset. Based on the text analysis table, several key findings emerge. First, public understanding of stunting remains varied, with the majority recognizing its physical impact (75.70%), but fewer being aware of its effects on intelligence (49.40%) and brain development (43.40%). Additionally, an information gap still exists, with 8.50% of respondents believing that stunting has no impact. Second, education plays a crucial role in raising public awareness, as evidenced by the strong connections of the word *edukasi* with terms such as further and primarily in the graph. This aligns with findings from a WHO study, which emphasized that intensive health campaigns and community-based education significantly enhance awareness and preventive actions against stunting **[28]**. Based on these findings, several recommendations can be proposed (**Table 5**):

a) Enhancing public education particularly on topics such as sanitation, poverty, and child growth monitoring, which remain underemphasized in public discussions.
b) Strengthening visual-based health campaigns utilizing infographics, pie charts, and bar charts, which have been proven effective in helping the public comprehend information more clearly.
c) Community-based intervention strategies where education is delivered through more interactive outreach programs, considering that one-third of respondents still hold misconceptions about stunting.

Therefore, by adopting a more systematic and data-driven approach, it is expected that information gaps can be minimized, and stunting prevention efforts can be carried out more effectively.

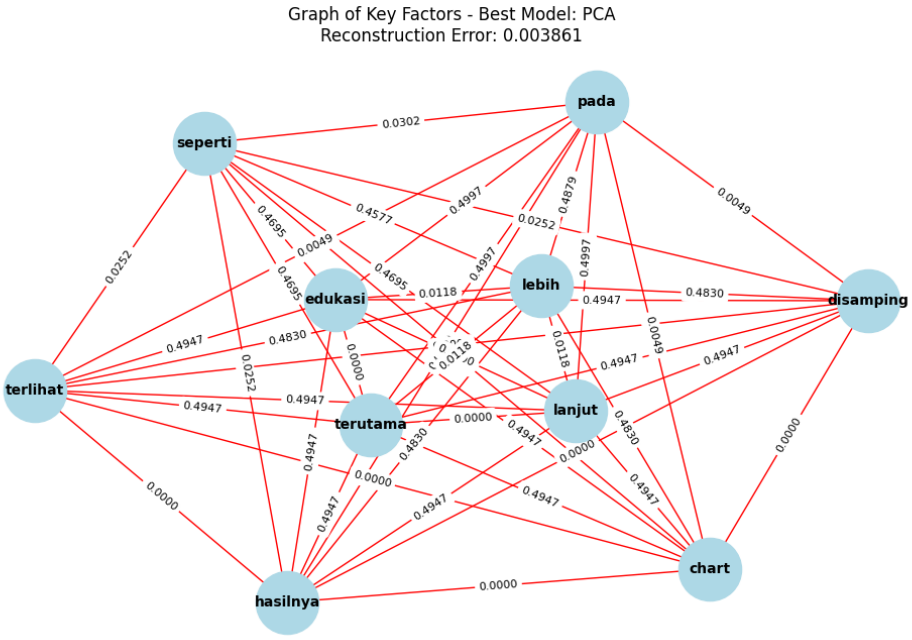

Figure 7. The Best Principal Factor Graph with Edge Weights from the Best Model

Table 5. Sentences Containing Main Factors

Index	Sentences
1	This shows that the majority of people have understood important aspects related to stunting, such as the causes, impacts and ways to prevent it.
2	However, the percentage who still have misconceptions is quite significant, almost a third of the total respondents.
3	The results are shown in the pie chart below.
4	Other factors, such as poor sanitation (27.90%) and hereditary diseases (25.10%), are still not given enough attention.
5	This data shows the need for further education, especially regarding sanitation and poverty, which also play a role in stunting prevention.
6	The results are shown in the bar chart below.
7	The SKI 2023 data shows that 75.70% of the community understands that stunting inhibits physical growth, while 49.40% realize its impact on intelligence and 43.40% on brain development.
8	More worryingly, 8.50% think stunting has no impact, indicating that there is still an information gap.
9	This data shows the need for further education, especially regarding sanitation and poverty, which also play a role in stunting prevention.
10	Further education is needed to increase public awareness, especially regarding child growth monitoring, immunization, and nutrition for pregnant women.
11	With more intensive health campaigns, it is hoped that the stunting rate can continue to be reduced.

Based on data analysis (**Table 5**), it was found that the majority of the population has understood key aspects related to stunting, such as its causes, impacts, and prevention methods. However, a significant percentage still holds misconceptions, indicating the urgent need for further education. Other factors, such as poor sanitation and hereditary diseases, remain largely overlooked by most respondents. This finding aligns with WHO, which emphasizes that sanitation and poverty are major contributing factors to the high prevalence of stunting in developing countries **[28]**. Additionally, the data reveal persistent information gaps regarding the impact of stunting on children's intelligence and brain development. Black et al explain that malnutrition during early life can have long-term consequences on cognitive development and overall health **[29]**. Therefore, more intensive health campaigns are necessary to enhance public awareness and reduce stunting rates, as recommended by UNICEF in its global strategy to promote nutritional awareness and education for families **[30]**.

This study is consistent with previous research, which indicates that public understanding of stunting remains variable. The study by UNICEF highlights the importance of continuous education to improve awareness of nutrition and maternal health as a preventive measure against stunting **[30]**. Another study by WHO also underscores the significant role of sanitation and poverty in stunting prevalence, reinforcing this study's findings regarding the need for intervention in these areas **[28]**. Furthermore, Dewey & Begum state that childhood stunting can lead to long-term cognitive impairment into adulthood **[31]**. Prendergast & Humphrey also emphasize that stunting prevention requires a multidisciplinary approach, including improved access to nutrition, sanitation, and healthcare services **[32]**. Thus, this study provides additional insight into the necessity of further education and a multidisciplinary approach in the prevention of stunting.

4. CONCLUSION

Based on the analysis results, the best-performing model is PCA, as it achieves a lower reconstruction error compared to SVD (0.003861 < 0.004232). The selection of the best model is based on the criterion of a smaller reconstruction error and its ability to identify key factors that significantly contribute to public understanding of stunting. PCA effectively highlights the strong relationships between key terms such as education, sanitation, and awareness of impacts, which are dominant factors in public discourse. While SVD captures more complex relationships, the simplicity and effectiveness of PCA in clustering key information make it the more optimal choice. These findings reinforce that data-driven approaches can enhance the effectiveness of educational campaigns and intervention strategies in stunting prevention efforts. However, this study is limited by the small sample size (21 reviews), which constrains the generalizability of the results. Future research should involve a larger and more diverse dataset to validate and refine the identified dominant factors. Additionally, integrating advanced text mining methods such as topic modeling (e.g., LDA) or neural network-based approaches could further uncover latent patterns and enhance the robustness of findings. Overall, the study demonstrates the value of applying data-driven, visual, and analytical methods to inform more targeted and effective public health communication strategies for stunting prevention.

REFERENCES

[1] S. Vijayarani, M. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining-an Overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[2] M. Ashtiani and B. Raahemi, "News-Based Intelligent Prediction of Financial Markets Using Text Mining and Machine Learning: A Systematic Literature Review," *Expert Syst Appl*, vol. 217, p. 119509, 2023, doi: 10.1016/j.eswa.2023.119509.

[3] D. Chrisinta and J. E. Simarmata, "Eksplorasi Teknik Web Scraping pada Data Mining: Pendekatan Pencarian Data Berbasis Python," *Faktor Exacta*, vol. 17, no. 1, pp. 1979–276, May 2024, doi: 10.30998/FAKTOREXACTA.V17I1.22393.

[4] D. Chrisinta and J. E. Simarmata, "Comparative Study of Support Vector Machine and Naive Bayes for Sentiment Analysis on Lecturer Performance," *Journal of Research in Mathematics Trends and Technology*, vol. 5, no. 1, pp. 1–7, 2023, doi: 10.32734/jormtt.v5i1.

[5] P. Monika, D. Devi Sri, and M. Suneetha, "Uncovering Insights in Agroforestry: A Text Mining Analysis Using Web Scraping and Topic Modeling," *Lecture Notes in Networks and Systems*, vol. 897, pp. 95–104, 2024, doi: 10.1007/978-981-99-9704-6_8.

[6] F. Nissa, A. Primandari, and A. Thalib, "Collaborative Filtering Approach: Skincare Product Recommendation Using Singular Value Decomposition (SVD)," *Media Statistika*, vol. 15, no. 2, pp. 139–150, 2023, doi: 10.14710/medstat.15.2.139-150.

[7] D. Delen and S. Ram, "Research Challenges and Opportunities in Business Analytics," *Journal of Business Analytics*, vol. 1, no. 1, pp. 2–12, Jan. 2018, doi: 10.1080/2573234X.2018.1507324.

[8] D. Freeman *et al.*, "COVID-19 Vaccine Hesitancy in The UK: the Oxford Coronavirus Explanations, Attitudes, and Narratives survey (Oceans) II," *Psychol Med*, vol. 52, no. 14, pp. 3127–3141, 2022, doi: 10.1017/S0033291720005188.

[9] P. Mitropoulos, K. Vasileiou, and I. Mitropoulos, "Understanding Quality and Satisfaction in Public Hospital Services: A Nationwide Inpatient Survey in Greece," *Journal of Retailing and Consumer Services*, vol. 40, pp. 270–275, 2018, doi: 10.1016/j.jretconser.2017.03.004.

[10] A. Fadllullah, D. D. Kamudi, M. Nasir, A. Zainal Arifin, and D. Purwitasari, "Web News Documents Clustering in Indonesian Language Using Singular Value Decomposition-principal Component Analysis (Svdpca) and Ant Algorithms," *Jurnal Ilmu Komputer dan Informasi*, vol. 9, no. 1, pp. 17–25, 2016, doi: 10.21609/jiki.v9i1.362.

[11] M. Greenacre *et al.*, "Principal Component Analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022, doi: 10.1038/s43586-022-00184-w.

[12] Y. Wang and L. Zhu, "Research and Implementation of SVD in Machine Learning," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, IEEE, 2017, pp. 471–475. doi: 10.1109/ICIS.2017.7960038.

[13] D. Chrisinta and J. E. Simarmata, "Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naïve Bayes Classifier," *Komputika: Jurnal Sistem Komputer*, vol. 12, no. 1, pp. 93–101, 2023, doi: 10.34010/KOMPUTIKA.V12I1.9638.

[14] M. Bala and D. Verma, "A Critical Review of Digital Marketing," *International Journal of Management, IT and Engineering*, vol. 8, no. 10, pp. 321–339, 2018.

[15] N. Rao, B. Naseeba, N. Challa, and S. Chakrvarthi, "Web Scraping (imdb) Using Python," *Telematique*, vol. 21, no. 1, pp. 235–247, 2022.

[16] S. Pradha, M. N. Halgamuge, and N. T. Q. Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," in *In 2019 11th international conference on knowledge and systems engineering (KSE)*, IEEE, 2019, pp. 1–8. doi: 10.1109/KSE.2019.8919368.

[17] S. Vijayarani and R. Janani, "Text Mining: Open Source Tokenization Tools-An Analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, 2016.

[18] D. Khyani, B. Siddhartha, N. Niveditha, and B. Divya, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," *Journal of University of Shanghai for Science and Technology*, vol. 22, no. 10, pp. 350–357, 2021.

[19] R. Kumbhar, S. Mhamane, H. Patil, S. Patil, and S. Kale, "Text Document Clustering Using K-Means Algorithm with Dimension Reduction Techniques," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 1222–1228. doi: 10.1109/ICCES48766.2020.9137928.

[20] A. Hassani, A. Iranmanesh, and N. Mansouri, "Text Mining Using Nonnegative Matrix Factorization and Latent Semantic Analysis," *Neural Comput Appl*, vol. 33, no. 20, pp. 13745–13766, Oct. 2021, doi: 10.1007/S00521-021-06014-6.

[21] Q. Wang and X. Cai, "Active-learning class activities and shiny applications for teaching support vector classifiers," *Journal of Statistics and Data Science Education*, vol. 32, no. 2, pp. 202–216, 2024, doi: 10.1080/26939169.2023.2231065.

[22] Y. Zhang, G. Li, and H. Zong, "A Method of Dimensionality Reduction by Selection of Components in Principal Component Analysis for Text Classification," *Filomat*, vol. 32, no. 5, pp. 1499–1506, 2018, doi: 10.2298/FIL1805499Z.

[23] K. Aris, C. Ramasamy, T. N. M. Aris, and M. Zolkepli, "Dynamic Force-directed Graph with Weighted Nodes for Scholar Network Visualization," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022, doi: 10.14569/IJACSA.2022.0131289.

[24] F. Nwokoma, J. Foreman, and C. Akujuobi, "Effective Data Reduction Using Discriminative Feature Selection Based on Principal Component Analysis," *Mach Learn Knowl Extr*, vol. 6, no. 2, pp. 789–799, 2024, doi: 10.3390/make6020037.

[25] Y. Melaku, G. Zello, T. Gill, R. Adams, and Z. Shi, "Prevalence and Factors Associated with Stunting and Thinness Among Adolescent Students in Northern Ethiopia: A Comparison to World Health Organization Standards," *Archives of Public Health*, vol. 73, no. 1, pp. 1–11, Oct. 2015, doi: 10.1186/s13690-015-0093-9.

[26] M. Marni *et al.*, "Cultural Communication Strategies of Behavioral Changes in Accelerating of Stunting Prevention: a Systematic Review," *Macedonian Journal of Medical Sciences*, vol. 9, no. F, pp. 447–452, 2021, doi: 10.3889/oamjms.2021.7019.

[27] S. Nanga *et al.*, "Review of Dimension Reduction Methods," *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, 2021, doi: 10.4236/jdaip.2021.93013.

[28] WHO, "World Malaria Report 2018," *J Phys A Math Theor*, vol. 44, no. 8, p. 085201, 2018, Accessed: Jan. 31, 2025. [Online]. Available: http://arxiv.org/abs/1011.1669

[29] R. E. Black *et al.*, "Maternal and child undernutrition and overweight in low-income and middle-income countries," *The Lancet*, vol. 382, no. 9890, pp. 427–451, 2013, doi: 10.1016/S0140-6736(13)60937-X.

[30] UNICEF, "Fed to fail? The crisis of children's diets in early life. United Nations Children's Fund.," Uny Press, 2021. Accessed: Mar. 31, 2025. [Online]. Available: https://www.unicef.org/reports/fed-to-fail-child-nutrition

[31] K. G. Dewey and K. Begum, "Long-term consequences of stunting in early life," *Matern Child Nutr*, vol. 7, no. Suppl 3, pp. 5–18, 2011, doi: 10.1111/j.1740-8709.2011.00349.x.

[32] A. J. Prendergast and J. H. Humphrey, "The stunting syndrome in developing countries," *Paediatr Int Child Health*, vol. 34, no. 4, pp. 250–265, 2014, doi: 10.1179/2046905514Y.0000000158.