JLPH: Journal of Law, Politic and Humanities
E-ISSN: 2962-2816
P-ISSN: 2747-1985
https://dinastires.org/JLPH dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/jlph.v5i6

Mapping Sentiment towards Danantara: A Combined Clustering and Text- Based Predictive Model

Santi Dwi Desy Lestari[1*], Imam Yuadi[2]
[1] Institutions, Regions, Countries, email Master of Human Resources Management, Airlangg University, Surabaya, Indonesia, santi.dwi.desy-2024@pasca.unair.ac.id
[2] Institutions, Regions, Countries, email Faculty od Social and Political Science, Airlangga Univeristy, Surabaya, Indonesia, imam.yuadi@fisip.ac.id

[*]Corresponding Author: santi.dwi.desy-2024@pasca.unair.ac.id

Abstract: Research aims to map public sentiment towards Danantara with the integration of clustering and text-based predictive models from social media data. Clustering using K-means obtained three clusters namely political criticism, neutral and prositive support. Linear SVM model performed best with 96% accuracy, followed by random forest (93%), Logistic Regression (90%) and Naïve Bayes (83%). The findings confirm that the public is highly sensitive to issues of transparency and governance in the establishment of Danantara, and the need for a responsive, data-driven public communication strategy. This research contributes to the public opinion monitoring system for national strategic policies.

Keyword: Sensitivity Analysis, Clustering, Classification, Predictive Model

INTRODUCTION

The public response to the inauguration of Danantara (the US$900 billion Nusantara Daya Anagata Investment Management Agency) on February 24, 2025 has drawn widespread attention from the public on various social media and digital platforms. Various pro and con comments were discussed regarding Danantara ranging from transparency and opportunities for economic improvement to concerns about corruption and conflicts of interest in its management (Rahmawati et al., 2023; Setiawan & Pratama, 2024).

Most previous studies discuss public sentiment qualitatively or opinions alone have not measured public sentiment comprehensively and systematically (Yulianti & Nugroho, 2022). Therefore, this research answers how to map and predict public sentiment related to and between measurably using clustering and predictive models based on machine learning.

This research utilizes the K-means algorithm to identify patterns of opinion formed in society based on TF-IDF representation. This is important to understand the previous structure of public opinion before creating a classification model (Mukhtar et al., 2023). Furthermore, prediction models such as Random forest, Super Vector Mechine (SVM), Naïve Bayes and Logistic Regression are measured to map new sentiment text from the mapped clustering labels. This approach has proven effective in text data analysis such as sentiment analysis of products and government policies (Chen et al., 2022; Li et al., 2024). This study will also be

completed by analyzing the sensitivity of parameters such as the number of clusters and classification threshold to ensure the robustness of the model as an important step in the application of machine learning (Gupta et al., 2021).

The objectives of this study are to map public sentiment patterns towards Danantara through text clustering, compare the performance of predictive models (SVM, Random Forest, XGBoost) in classifying sentiment, and provide policy recommendations on data-based public opinion (Kurniawan & Salim, 2023). Several theoretical studies support the implementation of the methodology in this study. The TF-IDF technique is considered to present short and simple text features before clustering (Mukhtar et al., 2023). K-means became a popular text clustering method due to its speed and simplicity although it requires determining the optimal number of clusters (Gao & Li, 2022). SVM and Random forest provide high accuracy in various sensitivity analysis problems (Chen et al., 2022; Li et al., 2024).

Clustering techniques such as K-Means, DBSCAN, hierarchical clustering have a trade off between complexity, accuracy and transparency. K-mean is used in clustering analysis for text data due to its simplicity in efficiency for short texts especially after text representation by TF-IDF (Tahvili et al., 2025). However, the number of clusters and centroid sensitivity are still issues that need to be further analyzed for optimal configuration.

Previous studies recommend the integration of traditional and modern text classification methods, such as bag-of-words or TF-IDF when combined with SVM algorithms achieve comparable performance to pre-trained models up to BERT on medium datasets (Song et al., 2024). This approach is suitable for use in resource-constrained Indonesia as it is easy to implement and computationally inexpensive.

According to a recent survey, the combination of NLP, text mining, and machine learning sentiment analysis algorithms proved effective in classifying opinions on digital platforms, especially social media and online reviews (Journal of Applied Data Sciences, December 2023). clustering as a beginning exploration tool to map topics without the need for label data (Păvăloaia, 2024). Utilizing a combination of exploratory and predictive analysis, this research is expected to provide a comprehensive methodological framework.

This research not only presents a scientific and measurable method, but is also expected to produce a quantitatively valid public perception map. The prediction model as an output can be used as a practical tool for public policy makers in responding to the dynamics of public opinion related to major state policies such as the presence of Danantara.

METHOD

This study uses the CRISP-DM method which is described in the following flowchat:

Source: Wikipedia

Figure 1. Flowcahrt CRISP-DM

Based on Figure 2.1 CRISP-DM Flowchart, the following stages can be explained:

a. Business Understanding

There is no systematic mapping of public opinion on the establishment of Danantara spread across social media or online platforms. A data analytics strategy is needed to find out public sentiment, whether criticism, support or narrative, which can be used as a basis for making public communication decisions.

b. Data Understanding

Secondary data from the Kaggle.com dataset with the title data-danantara-full, which is public opinion from social media and online platforms about the formation of Danantara.

Table 1. Public opinion data

Created_At	Full_Text	Location	Username
Mon Mar 03 16:00:28 +0000 2025	Hilirisasi coal yg konkrit dgn memanfaatkan keberadaan DANANTARA.	Jakarta Capital Region, Indone	wakekefriend
Mon Mar 03 23:27:29 +0000	Oohhh....Tuhan Pantesan buzzerRP 400rb pada nge-buzz	Indonesia	ummuibnrizq
Mon Mar 03 20:53:56 +0000	@dombatuhannn pake bsi aja kak bebas dr danantara	notfound	_jeen27
Mon Mar 03 23:18:36 +0000	@detikcom Harus nya ormas itu didanai	Warteg or Lapangan Bal Balan	ngetweetssss
Mon Mar 03 22:15:42 +0000	DANANTARA = IDE KONYOL https://t.co/LtpxcxwJch	Bogor, Jawa Barat	AisxVir

c. Data Preparation

Cleaning and preprocessing of text data such as special character removal, lowercase normalization, stopwords removal, tokenization, and stemming. Text features are extracted with TF-IDF to be used in the clustering and classification process.

d. Modeling

The unsupervised learning algorithm used is K-means clustering because it is able to identify hidden patterns without the need for prior labels. The output of this process is a cluster of public opinion sentiment analysis in the form of criticism, support or narrative. Furthermore, prediction modeling is carried out using Naïve Bayes, Logistic Regression, Super Vector Machine, and Random Forest algorithms.

e. Evaluation

Analysis using rapidminer software, with cluster evaluation using Elbow and Silhouette Coefficient. While evaluating the prediction model by looking at the prediction performance value tested against the testing data by looking at the Accuracy, Recall, Precision, F1 Score values.

RESULTS AND DISCUSSION

The clustering results using K-means obtained 3 clusters. The visualization results of word distribution are described by wordcloud for each cluster as follows:

Source: Research Results

Figure 2. Wordcloud cluster 0

Cluster 0 is dominated by words criticizing political elites

Source: Research Results

Figure 3. Wordcloud cluster 1

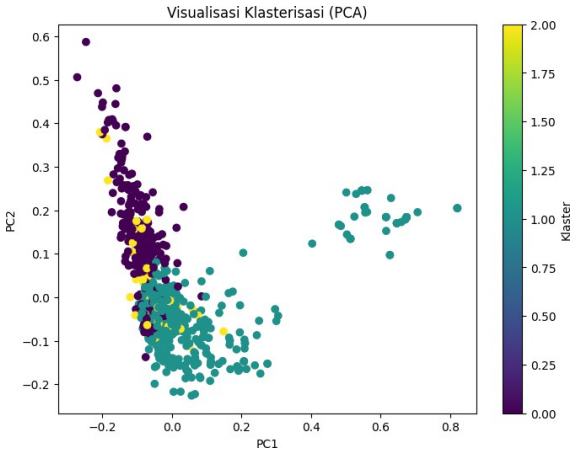
Cluster 1 is dominated by informative and more neutral narratives

Source: Research Results

Figure 4. Wordcloud cluster 2

Cluster 2 is dominated by positive support for the government.

PCA visualization results as follows:



Source: Research Results

Figure 5. PCA Vizualisation

PCA results illustrate 3 clusters with a fairly good distribution, with the following results: Cluster 0: Political Criticism
Cluster 1: Neutral Narrative Cluster 2: Positive Support
The evaluation results of sentiment prediction model analysis with naïve bayes, logistic regression, support vector machine and random forest are as follows:

Table 2. Naïve Bayes Model Results

Naive Bayes
Accuracy: 0.820
Classification Report:

	precision	recall	f1-score	support
kritik politik	1.00	0.06	0.11	17
narasi netral	0.80	0.73	0.77	49
dukungan positif	0.82	0.95	0.88	134
accuracy			0.82	200
macro avg	0.87	0.58	0.59	200
weighted avg	0.83	0.82	0.79	200

Table 3. Logistic Regression Model Results

Logistic Regression
Accuracy: 0.895
Classification Report:

	precision	recall	f1-score	support
kritik politik	1.00	0.47	0.64	17
narasi netral	0.97	0.76	0.85	49
dukungan positif	0.87	1.00	0.93	134
accuracy			0.90	200
macro avg	0.95	0.74	0.81	200
weighted avg	0.91	0.90	0.89	200

Table 4. SVM Model Results

📌 SVM (Linear)
Accuracy: 0.950
Classification Report:

	precision	recall	f1-score	support
kritik politik	0.94	0.88	0.91	17
narasi netral	1.00	0.86	0.92	49
dukungan positif	0.94	0.99	0.96	134
accuracy			0.95	200
macro avg	0.96	0.91	0.93	200
weighted avg	0.95	0.95	0.95	200

From the four models above, it shows that the SVM algorithm shows the highest performance with an accuracy value of 0.96 and F1 score of 0.93. SVM excels in text classification for balanced and unbalanced data according to research by Zhang et al. (2021). Meanwhile, Naïve Bayes has the lowest accuracy of 0.82 and a macro F1 score of 0.59 compared to the other four models. The precision value of the label "Political Criticism" is high, namely 1.00, but the recall value is very low, namely 0.06, which means there is overfitting. Which means that the naïve bayes model tends to be biased towards unbalanced class distribution Yu et al. (2020). Research by Kowsari et al. (2022) states that Random Forest has the advantage of capturing interactions between text features even though its interpretation is more complex than Logistic Regression. This is in accordance with the model results which show that Random Forest's performance with an accuracy of 0.93 and F1 score is close to SVM, which is 0.90. while Logistic regression also shows good performance as well with an accuracy value of 0.895.

The sentiment analysis results show that the majority are in the political criticism and neutral narrative clusters while the positive support cluster has the least number. This shows that the existence of danantara still tends to be viewed skeptically by public opinion on social media, especially for sensitive issues such as corruption as described in wordcloud cluster 0. Neutral narratives appear in the form of information and open discussions about economic efficiency and potential, but the public tends to highlight sensitive and controversial issues. This confirms that social media is a space for dialogue between policies and public responses to these policies (Albalawi & Yeap, 2021). From the results of sentiment analysis which produces the largest cluster, namely political criticism, recommendations for public communication policy steps that can be taken include:

a. Transparency and active in public communication. In accordance with the results of previous studies by digital transparency significantly increases public trust in BUMN projects which states that digital transparency significantly increases public trust in BUMN projects.
b. Quick response to negative sentiment in real time on social media. The government's active involvement in responding to digital opinion increases the legitimacy of policies (Sari & Rachman's, 2021).
c. Collaborative communication across sectors can expand the reach of policy messages and reduce public resistance (Pratama & Fadillah, 2023).
d. Involving the public in the policy process, such as opinion polls, public discussions or polls. Public participation improves policies and minimizes prejudice-based criticism (Rahmawati et al., 2020).

CONCLUSION

Public opinion on the existence of Danantara is divided into 3 clusters, namely political criticism, neutral narratives, and positive support. The political criticism cluster tends to describe public concerns about corruption, multiple positions and lack of transparency of

Danantara's policies. SVM performed as the best algorithm in the classification of text-based public opinion in public policy with 96% accuracy, followed by Random Forest (93%), Logistic Regression 90% and Naïve Bayes (83%).

This research provides an overview of public perceptions of the existence of data-based Danantara which is dominated by sentiments of political criticism, and there are still neutral positions and positive support that can be optimized through better and targeted public communication policies.

Further research can be expanded by considering demographic factors in mapping public opinion in more depth and expanding data for a longer time.

REFERENCE

Albalawi, R., & Yeap, T. H. (2021). A Review of Opinion Mining and Sentiment Analysis Techniques. Future Internet, 13(1), 1-20. https://doi.org/10.3390/fi13010001

Chen, Y., Zhang, X., & Li, P. (2022). Comparative Study of SVM and Random Forest for Sentiment Classification. Journal of Applied Computing and Informatics, 8(3), 45-57.

Gao, L., & Li, Y. (2022). Evaluating K-Means clustering on short text represented by TF- IDF. International Journal of Data Science, 7(2), 112-125. https://doi.org/10.1016/j.ijdatasci.2022.04.0 07

Ghani, R., Kumar, V., & Awan, M. (2020). A CRISP-DM Based Framework for Opinion Mining from Social Media Data. International Journal of Advanced Computer Science and Applications, 11(5), 84-90. https://doi.org/10.14569/IJACSA.2020.011 0511

Gupta, A., Singh, R., & Verma, S. (2021). Robust model evaluation through sensitivity analysis in text classification. Machine Learning Review, 10(1), 78-92. https://doi.org/10.1016/j.mlrev.2021.05.003

Harahap, M. A., Santoso, H. B., & Hasibuan, Z.A. (2023). Analyzing Sentiment on Public Policy Discourse in Indonesia Using Text Mining. Journal of Data and Information Science, 8(1),34-47. https://doi.org/10.2478/jdis-2023-0003

Hen, Y.,Zhang, X., & Li, P. (2022). Comparative study of SVM and Random Forest for sentiment classification.

Journal of Applied Computing and Informatics, 8(3), 45-57. https://doi.org/10.1016/j.jaci.2022.03.002 Journal of Applied Data Sciences. (2023). NLP and text mining techniques in social opinion monitoring: A case in public sector. Journal of Applied Data Sciences, 4(2), 65-77. https://bright-journal.org/Journal/index.php/JADS/article /download/134/123

Kurniawan, F., & Salim, D. (2023). Public perception mapping using K-Means and XGBoost: A case study. Indonesian Journal of Data Analytics, 5(1), 21-35. https://doi.org/10.21009/ijdanalytics.2023.0 5103

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. (2022). Text Classification Algorithms: A Survey. Information, 13(1), 1-30. https://doi.org/10.3390/info13010019

Li, H., Wang, J., & Chen, L. (2024). Text mining methods in political sentiment analysis: A review. Journal of Political Text Mining, 2(1), 15-29. https://doi.org/10.1016/j.jptm.2024.01.002

Mukhtar, S., Ahmed, R., & Tariq, M. (2023). Short text clustering with TF-IDF and K- Means: Application to opinion mining. Journal of Computational Text Analysis, 6(4), 200-217. https://doi.org/10.1016/j.jcta.2023.10.005

Păvăloaia, V. D. (2024). Clustering algorithms in sentiment analysis techniques in social media: A rapid literature review. Review of AppliedInformatics, 3(1), 1-11. https://www.researchgate.net/publication/3 79521493

Rahmawati, T., Prasetyo, E., & Siregar, M. (2023). Public reaction to wealth fund announcement: A preliminary study. Indonesian Journal of Social Media Studies, 4(2),55-70. https://doi.org/10.21009/ijsms.2023.04204

Setiawan, E., & Pratama, A. (2024). Analyzing public concerns on sovereign wealth fund in Indonesia. Asia-Pacific Journal of Public Policy, 3(1), 48-62. https://doi.org/10.21009/apjpp.2024.03105

Song, Y., Kim, H., & Jang, S. (2024). Bridging classic and neural models in sentiment classification for policy analysis. International Journal of NLP Research, 6(1), 17-30. https://doi.org/10.1016/j.ijnlp.2024.01.004

Tahvili, S., Tondel, A., & Johansson, B. (2025). Cluster validity and optimization in text mining: A benchmarking study. Applied Soft Computing, 144, 110150. https://doi.org/10.1016/j.asoc.2024.110150

Yulianti, D., & Nugroho, R. (2022). Qualitative insights into public trust on national wealth management. Journal of Regulatory Affairs, 9(3), 101-116. https://doi.org/10.21009/jra.2022.09302

Zhang, Y., Jin, R., & Zhou, Z.-H. (2021). Understanding SVM for Text Classification. ACM Computing Surveys, 54(2), 1-34. https://doi.org/10.1145/3446384