JENIUS
Scientific Journal, Human Resource Management
p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

Analyzing the #KaburDuluAja Phenomenon as an Indicator of Brain Drain in Indonesia: A Hybrid Text Mining Approach

Ulfa Meilinda Putri[1*], Hendro Margono[2]
Airlangga University, Surabaya, Indonesia
ulfameilinda@gmail.com[1*]

Received 26 March 2026 | Revised 30 March 2026 | Accepted 01 April 2026
* Correspondence Author

Abstract

In recent years, an increasing number of Indonesian youth have expressed a desire to seek better opportunities and life prospects abroad. One of the hashtags widely used to articulate this aspiration is #KaburDuluAja. *This hashtag has emerged as an expression of emotions, anxiety, and even frustration among certain segments of society, particularly the younger generation, in response to the current social, economic, and political conditions in Indonesia. This study aims to comprehensively examine the phenomenon of public discourse on social media related to the* #KaburDuluAja *hashtag as a reflection of the potential for brain drain in Indonesia. To achieve this objective, a hybrid approach is employed, integrating Clustering, Association Rule Mining, and Sentiment Classification methods. The research adopts a quantitative approach supported by text mining techniques and social media data analysis. Furthermore, this study is both descriptive and predictive in nature. The research does not focus on a specific geographical location; rather, it examines public conversations on social media platforms, particularly Twitter (currently known as X). The data utilized in this study were collected through a web crawling process. The findings indicate that public conversations related to the* #KaburDuluAja *phenomenon can be categorized into five main clusters. The evaluation of several classification models reveals that the Naïve Bayes algorithm achieves the highest accuracy in predicting sentiment, reaching 99.85%. The K-Nearest Neighbor (KNN) model achieves an accuracy of 75.79%, while the Decision Tree and Random Forest models demonstrate relatively lower performance, with accuracy levels around 60%. The* #KaburDuluAja *phenomenon is not merely a form of humor or casual expression on social media; rather, it represents a genuine reflection of social unrest, particularly concerning employment opportunities, welfare conditions, and the future prospects of Indonesia's younger generation.*

Keyword: *Association Rule Mining; Brain Drain; Clustering; Hybrid**; #KaburDuluAja; Sosial Media***

INTRODUCTION

In recent years, an increasing number of Indonesian youth have expressed a desire to seek better opportunities and life prospects abroad. This trend is reflected not only in the growing number of citizens who work, study, or migrate overseas, but also in its increasing prominence in social media discussions. One of the hashtags widely used to express this aspiration is *#KaburDuluAja*. According to an international research organization, YouGov, approximately 41% of Generation Z (born between 1997 and 2009) express a desire or are considering relocating abroad (tim CNN, 2025).

Although often framed in humor or sarcasm, the *#KaburDuluAja* phenomenon should not be taken lightly. Beneath this expression lies a genuine sense of concern regarding the domestic situation. Issues such as the difficulty of obtaining decent employment, low levels of welfare, social inequality, political uncertainty, and the suboptimal quality of public services are among the key factors that frequently drive the desire to "leave" Indonesia (World Bank, 2019); (International Labour Organization, 2022). This phenomenon is closely associated with what is commonly referred to as *brain drain*, namely the outflow of highly skilled human resources from Indonesia to other countries in pursuit of better living conditions and future prospects. If not addressed seriously, brain drain may lead to the loss of valuable human capital that should otherwise serve as a key driver of national development.

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

From a theoretical perspective, this study is grounded in contemporary developments of Push–Pull Migration Theory**,** which remains widely used in recent migration studies to explain the interaction between structural pressures in origin countries and opportunities in destination countries. Recent literature highlights that push factors such as unemployment, income inequality, and socio-political uncertainty continue to play a significant role in shaping migration intentions, while pull factors are increasingly associated with global labor market integration and digital exposure to opportunities abroad (Carling & Collins, 2018; Czaika & Reinprecht, 2022). In the Indonesian context, these dynamics are reflected in digital conversations that emphasize dissatisfaction with domestic conditions and aspirations for better opportunities overseas. On the other hand, social media has become a highly effective platform for capturing public views and sentiments on various issues, including migration intentions. Twitter, characterized by its speed, openness, and high level of interaction, has emerged as one of the primary platforms where discussions surrounding the *#KaburDuluAja* phenomenon take place. Through the analysis of conversations occurring on this social media platform, a more genuine and spontaneous understanding can be obtained regarding what people truly feel and think.

In addition, this study adopts Human Capital Theory from a modern perspective, which views migration as a strategic decision by individuals to maximize the returns on their education, skills, and competencies in a globalized economy. Recent studies suggest that highly educated individuals are more likely to engage in international mobility when domestic labor markets fail to absorb skilled workers effectively (Beine, Docquier, & Rapoport, 2022). This indicates that migration is not merely a reaction to unfavorable conditions, but also a proactive investment decision driven by rational economic considerations.

Furthermore, this research is informed by digital sociology, particularly the understanding that social media platforms function as spaces for constructing and amplifying public discourse. Recent studies emphasize that digital platforms such as Twitter (X) enable the formation of collective narratives that reflect societal concerns, emotions, and perceptions in real time (Vicari & Murru, 2020; Rogers, 2021). In this context, the hashtag *#KaburDuluAja* can be conceptualized as a form of digital migration discourse, where individual expressions aggregate into a broader social narrative regarding migration intentions and dissatisfaction with domestic conditions.

To comprehensively understand this phenomenon, it is insufficient to examine only a limited portion of the conversations. A comprehensive and systematic approach is required to process and analyze large-scale data. Therefore, this study integrates multiple analytical methods, referred to as a hybrid approach, in order to obtain a holistic and objective understanding of the *#KaburDuluAja* phenomenon. The first method employed is Clustering, which aims to group conversations or tweets into several main themes or topics (Sabna, et al., 2020). Through this method, it is possible to identify whether public discussions are predominantly focused on issues such as salary, employment, education, political conditions, or other factors that motivate individuals to seek opportunities abroad.

Despite the growing number of studies utilizing sentiment analysis and text mining techniques, several limitations persist in recent literature. Most prior studies tend to rely on single-method approaches**,** focusing primarily on classification accuracy without exploring the relationships between topics and underlying factors (Zhang, 2023). Other studies emphasize clustering techniques to identify discussion themes but lack integration with sentiment dynamics and relational analysis (Pavaloaia, 2024). Moreover, limited research has explicitly linked computational text analysis with migration theory, resulting in a gap between technical findings and theoretical interpretation. In addition, this study is complemented by the use of Association Rule Mining, which is employed to uncover patterns of relationships among words or terms that frequently co-occur within conversations (Lowin, 2024). This method can assist in identifying interrelated factors that are frequently mentioned together in discussions regarding migration intentions, such as the relationships among the terms “salary,” “employment,” “overseas,” and “future.”

Furthermore, Sentiment Analysis or Classification is employed to determine the overall tone or underlying sentiment of public conversations (Zhang, 2023). Through this analysis, it can be determined whether the conversations predominantly convey positive sentiments (optimism toward Indonesia’s conditions), negative sentiments (characterized by disappointment and dissatisfaction), or neutral sentiments (mere opinions without strong emotional judgment). By integrating these three methods, this study is expected to provide a more comprehensive and in-depth understanding of the dynamics of public opinion regarding the *#KaburDuluAja* phenomenon, including the key issues discussed, prevailing

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

sentiment tendencies, and the patterns of interrelationships among factors influencing migration intentions abroad.

Therefore, the novelty of this study lies in three key contributions. First, it proposes a hybrid analytical framework that integrates clustering, association rule mining, and sentiment analysis to capture the multidimensional structure of public discourse. Second, it bridges migration theory and digital data analytics, enabling a theoretically grounded interpretation of social media conversations related to migration intentions. Third, this study introduces the concept of **"**social media–driven brain drain perception,**"** which highlights how digital narratives can serve as early indicators of potential migration trends among the younger generation. By combining theoretical insight with computational methods, this research advances beyond descriptive analysis and offers a more analytical and integrative understanding of the *#KaburDuluAja* phenomenon.

The findings of this study are not only valuable for understanding migration phenomena from a societal perspective, but also serve as important input for policymakers, academics, and other stakeholders in formulating more targeted policies to mitigate brain drain, enhance national competitiveness, and foster improved social, economic, and political conditions. Ultimately, such efforts are expected to encourage Indonesian society, particularly the younger generation, to feel more secure and optimistic about remaining and contributing to their home country.

METHOD

This study employs a quantitative method supported by text mining techniques and social media data analysis. The quantitative approach is selected as the study focuses on the collection, processing, and analysis of large-scale textual data, which are subsequently transformed into numerical information that can be statistically interpreted to address the research problem (Sugiyono, 2020). This study is also characterized as both descriptive and predictive in nature. The descriptive aspect aims to provide a detailed overview of conversation clustering patterns, public sentiment tendencies, and the relationships among words or terms that frequently appear in discussions related to the phenomenon. The predictive aspect is implemented through the application of classification models to predict the sentiment of public conversations based on previously labeled data.

The object of this study is the conversational data of Indonesian users on Twitter involving the *#KaburDuluAja* hashtag, which is considered to represent concerns, aspirations, or the desire to seek better opportunities abroad. The population in this study comprises all public conversations on Twitter containing the *#KaburDuluAja* hashtag or keywords related to migration intentions and the potential for brain drain. The sampling technique employed is non-probability sampling, specifically purposive sampling, in which samples are selected based on predetermined criteria (Sugiyono, 2020). The criteria applied in this study require that the tweets must contain the *#KaburDuluAja* hashtag or relevant keywords related to migration intentions or conditions in Indonesia, and must be written in the Indonesian language. Only publicly accessible and legally available tweets are included as research data. The sample size is determined based on the capacity of the tools used and the analytical requirements in order to obtain representative results.

The data collection method employed to enhance the validity of this study is web crawling, which is an automated data retrieval technique from the Twitter (X) social media platform using the *#KaburDuluAja* keyword. Data collection was conducted using Google Colab tools integrated with Python libraries such as Tweepy or snscrape, enabling efficient large-scale data extraction. The types of data collected include tweet text or conversation content, timestamps (date and time of posting), number of likes, number of retweets, and other relevant information for analytical purposes. In addition to the primary data in the form of tweet text, this study also collects supplementary information in the form of related keywords, such as "work," "overseas," "salary," "education," "Indonesia," and other relevant terms.

The data analysis procedure in this study consists of several stages. First, data preprocessing is conducted to ensure that the dataset is clean, relevant, and ready for analysis. This process is performed using RapidMiner. The preprocessing steps include removing irrelevant elements such as symbols, numbers, URLs, emoticons, and punctuation; converting all text into lowercase (case folding); removing stopwords based on a combined list of general stopwords and customized stopwords derived from preliminary observations; performing tokenization to split sentences into individual words; and applying

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

stemming to transform inflected or derived words into their base forms. Second, clustering analysis is applied to group tweets into several main topics related to the *#KaburDuluAja* phenomenon. The clustering method used is k-Means clustering, with the number of clusters determined based on initial exploratory analysis (Regina, et al., 2021). This stage aims to identify the dominant issues discussed by the public, such as topics related to employment, education, economic conditions, and dissatisfaction with government policies. Third, Association Rule Mining (ARM) is conducted using the FP-Growth algorithm to identify patterns of relationships among words or terms that frequently co-occur in the conversations (Rasianto, R., & Sutedi, S.,2023). This analysis aims to uncover combinations of words or phrases that consistently appear together, thereby providing insights into interrelated factors influencing migration intentions. The results of this analysis are presented in the form of association rules, measured by support, confidence, and lift values, which indicate the strength of relationships between terms. Fourth, sentiment analysis is performed to determine the emotional tendency of public conversations related to the *#KaburDuluAja* phenomenon. The sentiment is categorized into three classes: positive (tweets expressing optimism or positive expectations toward Indonesia), negative (tweets reflecting pessimism, dissatisfaction, or strong intentions to migrate), and neutral (tweets that are informative, express neutral opinions, or lack emotional judgment). Fifth, the manually labeled dataset is divided into training and testing sets to develop sentiment classification models using several machine learning algorithms, including Naïve Bayes, Decision Tree, and Random Forest (Saputra et al., 2023). Finally, model evaluation is conducted to assess the performance of the classification models. Several evaluation metrics are employed, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of model effectiveness.

RESULT AND DISCUSSION

Result

Clustering analysis

Clustering analysis was conducted to group public conversations on Twitter related to the *#KaburDuluAja* phenomenon into several clusters or main themes based on the similarity of their content. The results of the clustering process identified five primary clusters, each with a different proportion, as presented in the following table.

Table 1 Result of Clustering

No	Cluster	Number of Conversation	Persentase
1	Cluster 0	1.857	36,4%
2	Cluster 3	1.730	33,9%
3	Cluster 1	671	13,2%
4	Cluster 2	637	12,5%
5	Cluster 4	205	4,0%

Based on these results, it can be concluded that Cluster 0 represents the largest group, accounting for approximately 36.4% of the total analyzed conversations. This indicates that the themes or topics within Cluster 0 constitute the most dominant issues discussed by the public regarding the *#KaburDuluAja* phenomenon. The next largest cluster is Cluster 3, comprising 33.9% of the conversations. Meanwhile, Cluster 1 accounts for 13.2%, Cluster 2 for 12.5%, and Cluster 4 represents the smallest group with only 4.0% of the total conversations.

Although each cluster is automatically labeled using numerical identifiers, a deeper interpretation requires further exploration of the content within each cluster. Based on preliminary analysis, the differences among clusters indicate the presence of thematic variations in discussions related to the phenomenon of migration abroad. These clustering results provide an initial overview of the structure of public discourse related to the *#KaburDuluAja* phenomenon. By organizing conversations into several main themes, subsequent analyses such as Association Rule Mining and Sentiment Analysis can be conducted more effectively, thereby enabling a more comprehensive understanding of the issues and public perceptions.

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

Association Rule Mining Analysis

Table 2 Result of Association Rule Mining

No.	Word/Phrase Combination	Support (%)	Interpretation
1	work → abroad	26,5%	The desire to work abroad emerges as a primary motive in the conversations
2	work → overseas	26,8%	The dominant topic among the public is seeking job opportunities abroad.
3	work → abroad → overseas	25,1%	The combination of these three terms indicates a strong intention to migrate for employment.
4	work → salary	11,6%	Low salary is identified as one of the driving factors for migration intentions.
5	work → difficult	18%	Difficulty in obtaining employment triggers the desire to move abroad.
6	graduate → degree holder → unemployed	10% - 11,5%	Unemployment among university graduates is identified as a significant issue.
7	abroad → overseas → better	10,1%	There is a positive perception that life abroad is more favorable.

Based on public conversations on social media using the *#KaburDuluAja* hashtag, strong and consistent patterns of word associations were identified, providing a comprehensive understanding of the relationships among issues or factors driving migration intentions, particularly among Indonesian youth. These association patterns not only reflect the co-occurrence frequency of terms but also represent the social construction of public perceptions regarding labor market conditions, economic realities, and future prospects in Indonesia. First, the associations among the terms "work," "abroad," and "overseas" are particularly prominent. The combination of "work" and "abroad" appears in 26.5% of conversations, while "work" and "overseas" co-occur in 26.8%. Furthermore, the simultaneous occurrence of all three terms—"work," "abroad," and "overseas"—is observed in 25.1% of the conversations. This pattern indicates that the aspiration to work abroad constitutes one of the primary motivations underlying the *#KaburDuluAja* narrative. It suggests that a significant proportion of individuals are not merely expressing a desire to leave Indonesia, but are strongly motivated to pursue better employment opportunities overseas.

Second, the relationships between "work" and "salary" (11.6%), as well as "work" and "difficulty" (18%), reflect public concerns regarding domestic labor conditions. The difficulty in obtaining decent employment, coupled with relatively low income levels, emerges as a significant push factor in shaping migration intentions. Third, the association pattern among the terms "graduate," "degree holder," and "unemployed" (10%–11.5%) highlights the challenges faced by university graduates in securing employment that meets their expectations. This pattern reveals that high unemployment among graduates is not only a structural issue but also contributes to dissatisfaction, which in turn drives the intention to migrate. Fourth, the combination of "abroad," "overseas," and "better" (10.1%) illustrates a positive perception of life abroad as more promising, comfortable, and advantageous. This reflects the presence of strong pull factors influencing migration aspirations. Overall, these patterns suggest that the *#KaburDuluAja* phenomenon is not merely a form of humor or a temporary emotional expression, but rather a tangible representation of interconnected social concerns. In this regard, the Association Rule Mining method proves to be effective in uncovering relationships among terms within public discourse, thereby providing deeper insights into the key issues underlying migration intentions among Indonesian youth.

Public Sentiment Trends

Based on Figure 1, negative sentiment dominates the conversations, with 8,277 tweets, accounting for approximately 55% of the total analyzed data, indicating a negative tendency. This finding suggests that more than half of public discussions related to this phenomenon are characterized by expressions of disappointment, dissatisfaction, concern, or criticism toward conditions in Indonesia, including social, economic, political aspects, and future prospects. Neutral sentiment comprises 6,116 tweets, or approximately 41% of the total conversations. This indicates that a substantial portion of users discuss the phenomenon in an informative manner, through humor, or by expressing opinions without strong emotional judgment, whether positive or negative. Positive sentiment is observed in only

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

616 tweets, representing approximately 4% of the data. This suggests that only a small proportion of the public expresses optimism, motivation, or appreciation regarding issues related to migration abroad.

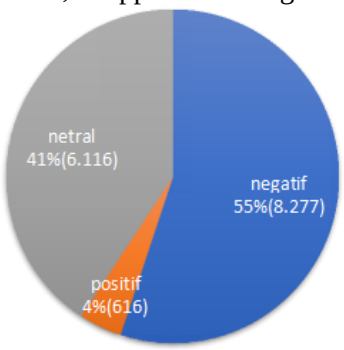

Figure 1. Sentiment Distribution of the #KaburDuluAja Phenomenon

The dominance of negative sentiment in conversations related to *#KaburDuluAja* indicates a relatively high level of concern and dissatisfaction among the public, particularly among the younger generation, regarding their current conditions. Issues such as the difficulty of obtaining decent employment, low levels of welfare, and the perception that life abroad offers better prospects serve as the main driving factors behind the emergence of this phenomenon. Meanwhile, the high proportion of neutral sentiment suggests that this phenomenon has also become a widely discussed topic; however, not all conversations exhibit a clear emotional tendency. The low proportion of positive sentiment further reinforces the assumption that the narrative surrounding migration abroad in this context is driven more by dissatisfaction than by optimism or appreciation of domestic conditions.

Classification Model

The initial stage involves data splitting, in which the dataset is divided into two parts: data used for model training and data used for model testing. The division is conducted using a 70:30 ratio for training and testing, respectively. Based on Figure 2, Negative sentiment represents the most dominant category, with approximately 2,763 data points. This indicates that the majority of public conversations related to the *#KaburDuluAja* phenomenon tend to carry a negative tone. In other words, many individuals express concerns, complaints, or dissatisfaction with the conditions they are currently experiencing. Neutral sentiment ranks second, with approximately 1,591 data points. This category generally consists of informative discussions, general opinions, or statements without strong emotional content, whether positive or negative. Positive sentiment constitutes the smallest proportion, with approximately 152 data points. This suggests that conversations reflecting hope, optimism, or appreciation regarding Indonesia's situation in the context of migration remain relatively limited.

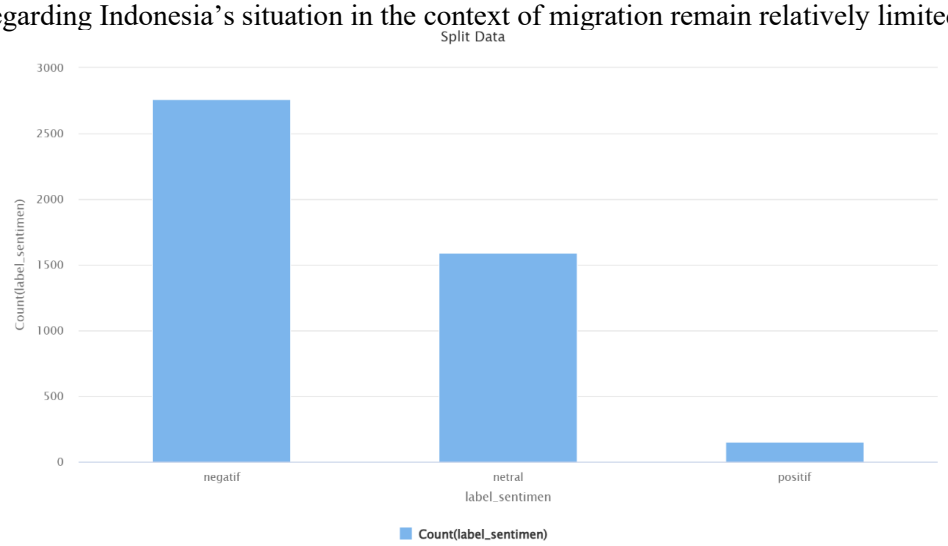

Figure 2. Result of Testing Data Sentiment Labels

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

Naïve Bayes Classification Model Evaluation

The Naïve Bayes model demonstrates the capability to predict public sentiment tendencies related to the *#KaburDuluAja* phenomenon, based on data that have undergone manual labeling and subsequent data splitting into training and testing sets. The performance of the model is evaluated using a confusion matrix, which reflects the comparison between the predicted labels generated by the model and the actual labels in the testing dataset.

Table 3 Confusion Matrix Results of the Naïve Bayes Model

Prediction/Real	True Negatif	True Netral	True Positif	Class Precision
Negatif Prediction	2.759	0	2	99,93%
Netral Prediction	4	1.591	0	99,75%
Positif Prediction	0	0	150	100,00%
Recall / Class	99,86%	100,00%	98,68%	-

Based on the results above, it can be observed that the Naïve Bayes model is able to predict negative sentiment with a very high level of accuracy. The overall accuracy of the model reaches 99.87%, indicating excellent performance. Out of the total testing data labeled as negative, 2,759 instances are correctly classified as negative. Only 4 negative instances are misclassified as neutral, and 2 instances are incorrectly predicted as positive. This indicates that the model demonstrates a very high level of sensitivity (recall) for the negative category, reaching 99.86%, with a precision of 99.93%, meaning that nearly all predictions classified as negative are indeed correct. For the neutral sentiment category, the Naïve Bayes model exhibits near-perfect performance. All 1,591 testing instances labeled as neutral are correctly predicted without any misclassification into other categories. The precision for the neutral category reaches 99.75%, while the recall achieves 100%, indicating the model's strong capability in accurately identifying neutral data.

Meanwhile, for the positive category, the model also demonstrates very strong performance, despite the relatively smaller number of instances. A total of 150 positive instances are correctly classified, with no misclassification into the neutral category; however, 2 instances are incorrectly classified as negative. The precision for the positive category reaches 100%, while the recall is 98.68%, indicating a very high level of accuracy and sensitivity in identifying positive sentiment.

Decision Tree Classification

Table 4. Confusion Matrix Results of the Decision Tree Model

Prediction/Real	True Negatif	True Netral	True Positif	Class Precision
Negatif Prediction	2.763	1.591	152	61,32%
Netral Prediction	0	0	0	0,00%
Positif Prediction	0	0	0	0,00%
Recall / Class	100,00%	0,00%	0,00%	-

Based on the results above, it can be observed that the total accuracy achieved by the Decision Tree model is 61.32%, indicating a significant decline compared to the previously evaluated Naïve Bayes model. The Decision Tree model tends to classify all data into the negative category, without generating predictions for the neutral or positive classes. Overall, although the accuracy of the Decision Tree model appears relatively high at 61.32%, this does not reflect good performance, as the model exhibits a bias toward the majority class (negative) while neglecting the minority classes (neutral and positive).

Table 5 *Confusion Matrix* Model *Random Forest*

Prediction/Real	True Negatif	True Netral	True Positif	Class Precision	Prediction/Real
Negatif Prediction	**Prediksi Negatif**	2.763	1.591	152	61,78%
Netral Prediction	**Prediksi Netral**	0	21	0	1,32%
Positif Prediction	**Prediksi Positif**	0	0	0	0,00%
Recall / Class	**Recall per Kelas**	100,00%	0,00%	0,00%	-

Based on the confusion matrix presented above, several conclusions can be drawn regarding the performance of the Random Forest model. The model exhibits a strong tendency to classify the majority of instances into the negative category. All instances originally labeled as negative are correctly predicted, resulting in a recall of 100% for the negative class. Compared to the Decision Tree model, Random Forest shows a slight improvement by correctly classifying a small portion of neutral instances, specifically 21 out of the total neutral data,

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

yielding a recall of 1.32% for the neutral category. However, the majority of neutral instances are still misclassified as negative. For the positive category, the model completely fails to recognize positive instances, as all positive data are predicted as negative, resulting in both recall and precision values of 0%. Although the overall model accuracy reaches 61.78%, this figure can be misleading, as it is largely influenced by class imbalance, where the proportion of negative data significantly exceeds that of neutral and positive classes. Consequently, the model tends to be biased toward predicting the majority class.

Table 6 *Confusion Matrix* Model K-Nearest Neighbor (KNN)

Prediction/Real	True Negatif	True Netral	True Positif	Class Precision	Prediction/ Real
Negatif Prediction	**Prediksi Negatif**	2.400	588	80	78,23%
Netral Prediction	**Prediksi Netral**	359	999	56	70,65%
Positif Prediction	**Prediksi Positif**	4	4	16	66,67%
Recall / Class	**Recall per Kelas**	86,86%	62,79%	10,53%	-

Based on the results above, it can be observed that the total accuracy of the KNN model is 75.79%. The K-Nearest Neighbor (KNN) model demonstrates relatively good performance in predicting the negative category. Out of a total of 2,763 data points that actually belong to the negative category, 2,400 were correctly classified, resulting in a recall value of 86.86% and a precision value of 78.23%. For the neutral category, the model demonstrates relatively adequate performance, correctly classifying 999 out of 1,591 actual neutral instances. This results in a recall of 62.79% and a precision of 70.65%, indicating that the model is more capable of identifying neutral sentiment compared to previous models such as Decision Tree and Random Forest.

However, the KNN model still faces challenges in identifying the positive category. Out of 152 actual positive instances, only 16 are correctly predicted, resulting in a very low recall of 10.53%, although the precision for the positive category is relatively high at 66.67%. This condition indicates that a substantial proportion of positive sentiment data is still misclassified by the model. Overall, the KNN model achieves an accuracy of 75.79%, which is relatively higher than that of Decision Tree and Random Forest, yet still falls short of the near-perfect performance demonstrated by the Naïve Bayes model. These findings suggest that while KNN has considerable potential, further optimization is required, particularly to improve prediction accuracy for the positive class.

Best Classification Model

Based on the evaluation results of several classification models, namely Naïve Bayes, Decision Tree, Random Forest, and K-Nearest Neighbor (KNN), it can be concluded that the Naïve Bayes model performs as the most effective model in predicting public sentiment related to the *#KaburDuluAja* phenomenon. The Naïve Bayes model demonstrates the most optimal performance, achieving an accuracy of 99.87%, along with high precision across classes and stable recall values for all sentiment categories (negative, neutral, and positive). This indicates that the model is highly capable of accurately classifying data with minimal prediction errors. In contrast, the Decision Tree and Random Forest models exhibit significantly lower accuracy, at approximately 61%, with major limitations in accurately identifying neutral and positive sentiment categories. Meanwhile, the KNN model performs better than Decision Tree and Random Forest, achieving an accuracy of approximately 75.79%; however, it still does not match the performance of Naïve Bayes, particularly in predicting the positive sentiment category.

The superior performance of the Naïve Bayes model in this study is supported by its simplicity and effectiveness in handling textual data, including preprocessed and standardized datasets. Moreover, this method is known to be relatively robust in handling class imbalance, which is evident in this study, where negative sentiment data significantly outnumber other categories. Therefore, it can be concluded that Naïve Bayes is the most suitable and reliable model for this study in mapping public sentiment tendencies toward migration phenomena, as expressed in social media conversations using the *#KaburDuluAja* hashtag.

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

Combined Analysis of Clustering, Association Rule Mining, and Sentiment Classification

The integration of Clustering, Association Rule Mining, and Sentiment Classification methods in this study provides a comprehensive understanding of the issues, public perceptions, and factors driving migration intentions among Indonesian youth on social media through the *#KaburDuluAja* hashtag.

First, through the application of clustering techniques, public conversations are successfully grouped into several clusters representing the main topics of public concern.

Subsequently, Association Rule Mining is employed to identify patterns of relationships among words or terms that frequently co-occur within the conversations. The analysis reveals strong associations among terms such as "work," "abroad," and "overseas," indicating a strong aspiration among the younger generation to work or reside abroad. In addition, associations among terms such as "work," "salary," and "difficulty" highlight the challenges of obtaining decent employment and adequate income in Indonesia. Furthermore, relationships among terms such as "graduate," "degree holder," and "unemployed" reflect concerns among educated individuals who continue to face unemployment. These association patterns demonstrate that migration intentions are not formed spontaneously, but are closely linked to complex and interrelated socio-economic conditions.

Meanwhile, sentiment analysis provides insights into the emotional perceptions of the public regarding this phenomenon. Based on manual labeling of the dataset, negative sentiment is found to dominate public conversations, indicating a high level of dissatisfaction, disappointment, and even frustration with domestic conditions. Neutral sentiment is also observed in a substantial proportion, reflecting informative or descriptive discussions without strong emotional expression. In contrast, positive sentiment appears in significantly smaller proportions, suggesting relatively low levels of public optimism toward the situation in Indonesia.

Overall, the integration of these three methods offers complementary insights into the *#KaburDuluAja* phenomenon. Clustering identifies the main topics of discussion, Association Rule Mining reveals the interrelationships among factors influencing migration intentions, and Sentiment Classification captures the emotional tendencies of public perception. Thus, this study not only explains what is being discussed but also elucidates how these issues are interconnected and how public perceptions and emotions are formed. These findings provide important insights for policymakers in formulating strategic interventions to mitigate brain drain and enhance the confidence of the younger generation in the future of Indonesia.

Discussion

The combination of Clustering, Association Rule Mining, and Sentiment Classification methods in this study provides a comprehensive understanding of the issues, public perceptions, and factors driving migration intentions among Indonesian youth on social media through the *#KaburDuluAja* hashtag. First, through the application of the Clustering method, public conversations were successfully grouped into several clusters representing the main topics of public concern. This finding is consistent with the study by Pavaloaia (2024), which states that clustering algorithms are effective in categorizing social media data into specific topic groups, thereby facilitating the identification of dominant issues (Pavaloaia, 2024). In addition, research also shows that a clustering based approach s able to dentify topic structures in social media data more systematically (Hanny and Resch, 2024).

The findings of this study indicate that public discourse surrounding the *#KaburDuluAja* phenomenon is dominated by negative sentiment, which reflects dissatisfaction with domestic socio-economic conditions. This result is consistent with recent studies showing that social media sentiment can serve as a proxy for public perception toward structural issues. For instance, a large-scale study on migration discourse across social media platforms found that negative sentiment tends to increase in response to socio-political pressures and critical events, indicating that online narratives are closely linked to real-world structural conditions (Nguyen, et al., 2026). This suggests that the dominance of negative sentiment in this study is not merely incidental, but reflects deeper systemic concerns such as employment insecurity and perceived inequality.

Furthermore, the strong association patterns identified between terms such as "work," "abroad," and "salary" reinforce the argument that economic motivations remain the primary drivers of migration intention. This finding aligns with recent computational social science research, which demonstrates that

sentiment and topic relationships extracted from social media data can reveal underlying socio-economic factors influencing public behavior and decision-making (Oller & Spadavecchia, 2024). In this context, the *#KaburDuluAja* discourse can be interpreted as a digital representation of push factors, particularly related to labor market constraints and limited economic opportunities.

In terms of methodology, the use of a hybrid approach combining clustering, association rule mining, and sentiment analysis provides a more comprehensive understanding compared to single-method studies. Previous research has highlighted that sentiment analysis alone often fails to capture the multidimensional structure of public discourse, particularly the relationships between topics and contextual factors (Hill, et al, 2025). Similarly, systematic reviews of sentiment analysis research emphasize the importance of integrating multiple analytical techniques to uncover not only sentiment polarity but also topic dynamics and relational patterns within large-scale textual data (Ibanez, et al, 2023). Therefore, the hybrid approach employed in this study represents a methodological advancement by enabling a deeper exploration of both thematic structures and sentiment tendencies.

However, the dominance of negative sentiment also requires a more critical interpretation beyond surface-level reasoning. From a socio-economic perspective, negative sentiment may reflect perceived relative deprivation, where individuals compare their current conditions with better opportunities abroad. This is supported by recent studies indicating that media and digital discourse surrounding migration are strongly influenced by socio-economic inequalities and perceived disparities in quality of life (Oller and Spadavecchia, 2024). Thus, the findings of this study highlight the role of social media as a space where structural dissatisfaction is collectively articulated and amplified.

Additionally, it is important to consider the limitations related to data representativeness. Prior studies emphasize that Twitter users do not fully represent the general population, as platform demographics tend to be skewed toward younger and more digitally active individuals (Nguyen, et al., 2026). This implies that the findings of this study primarily reflect the perceptions of digitally engaged youth, who are also the demographic most likely to consider international migration. Therefore, while the results provide valuable insights, they should be interpreted with caution in terms of generalizability.

Second, the Association Rule Mining method was employed to uncover patterns of relationships among words or terms that frequently co-occur in the conversations. The analysis results indicate strong associations among terms such as “work,” “abroad,” and “overseas,” reflecting a strong intention among the younger generation to work or reside abroad. In addition, associations among the terms “work,” “salary,” and “difficulty” were identified, indicating the challenges of obtaining decent employment and adequate income in Indonesia. Furthermore, the relationships among terms such as “graduate,” “degree holder,” and “unemployed” reflect the concerns of educated individuals who continue to face unemployment. Third, negative sentiment dominates public conversations, indicating a high level of dissatisfaction, disappointment, and even frustration regarding domestic conditions. Positive sentiment is found in significantly smaller proportions, suggesting that public optimism toward the situation in Indonesia remains relatively low. This study is consistent with prior research indicating that sentiment analysis of social media can effectively reveal public perceptions and emotions, and serves as an important indicator in understanding public opinion toward a particular phenomenon (Nip and Berthelier, 2024).

Overall, the combination of these three methods is complementary in providing a comprehensive overview of the #KaburDuluAja phenomenon. Clustering reveals the main topics of conversation, Association Rule Mining maps the interconnections between factors influencing migration intentions, while Sentiment Classification illustrates the public's emotional perception. Thus, this research not only explains what the public is discussing but also clarifies how these issues are interrelated and how public perceptions and emotions are formed, providing vital input for formulating strategic policies to curb brain drain and increase the younger generation's confidence in Indonesia's future.

CONCLUSSION

Based on the results of this study examining the *#KaburDuluAja* phenomenon on social media as a reflection of potential brain drain in Indonesia using a hybrid approach, it is found that public conversations related to this phenomenon can be categorized into five main clusters. Through the application of Association Rule Mining, strong patterns of word associations were identified, particularly among terms such as “work,” “abroad,” “salary,” “difficulty,” and “graduate.” These patterns reflect the key issues driving migration intentions, especially among the younger generation,

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

including limited employment opportunities, low income levels, and uncertainty regarding future prospects.

Furthermore, based on manual labeling results, negative sentiment dominates the conversations related to *#KaburDuluAja*. The majority of users express feelings of disappointment, pessimism, and even frustration toward the social, economic, and political conditions in Indonesia. Positive sentiment appears only in a small proportion, primarily in the form of hope or motivational expressions to persevere despite existing challenges.In terms of classification performance, the Naïve Bayes model demonstrates the highest accuracy in predicting sentiment, achieving 99.85%. The K-Nearest Neighbor (KNN) model achieves an accuracy of 75.79%, while the Decision Tree and Random Forest models show relatively lower performance, with accuracy levels of approximately 60%. The integration of these three analytical methods successfully provides a comprehensive understanding of public perception and the factors influencing migration intentions.

Overall, the findings suggest that the *#KaburDuluAja* phenomenon is not merely a form of humor on social media, but rather a tangible reflection of social unrest, particularly in relation to employment conditions, welfare, and the future of Indonesia's younger generation. Therefore, serious efforts are required to improve the quality and accessibility of employment opportunities, especially for university graduates. Enhancing labor market conditions, increasing fair wages, and ensuring greater certainty of future prospects within the country are essential steps to mitigate the risk of brain drain.
For future research, it is recommended to utilize larger and more diverse datasets from multiple social media platforms, as well as to consider the application of data balancing techniques in order to improve the robustness and accuracy of classification models.

Acknowledge

The author would like to express deepest gratitude to the supervisor for providingguidance andinput during the research process. Appreciation is also extended to the university for providing the opportunity, allowing this research process to proceed smoothly. Most importantly, special thanks are due to the author's parents for their endless prayers and support,which made this research possible.

References

[1] Beine, M., Docquier, F., & Rapoport, H. (2022). Brain drain and human capital formation in developing countries: Winners and losers. *The Economic Journal, 132*(646), 1903–1940. https://doi.org/10.1093/ej/ueac010

[2] Carling, J., & Collins, F. (2018). Aspiration, desire and drivers of migration. *Journal of Ethnic and Migration Studies, 44*(6), 909–926. https://doi.org/10.1080/1369183X.2017.1384134

[3] Czaika, M., & Reinprecht, C. (2022). Migration drivers in a digital age: Understanding aspirations and capabilities. *Journal of Ethnic and Migration Studies, 48*(12), 2835–2852. https://doi.org/10.1080/1369183X.2021.1986267

[4] Hanny, D., & Resch, B. (2024). Clustering-based joint topic-sentiment modeling of social media data: A neural networks approach. *Information (Switzerland), 15*(4), 200.

[5] Hill, C., Irshadiat, F., Johnson, M., & Fresneda, J. (2025). An analytical assessment of sentiment analysis trends and methods through systematic review and topic modeling. *Decision Analytics Journal, 17*. doi:https://doi.org/10.1016/j.dajour.2025.100644

[6] Ibanez, M. R., Ventura, A. C., Mateos, F. C., & Jimenez, P. M. (2023). A review on sentiment analysis from social media platforms . *Expert Systems With Applications* , 223.

[7] International Labour Organization. (2022). *World employment and social outlook: Trends 2022.* Geneva: ILO.

[8] Lowin, M. (2024). A Text-Based Predictive Maintenance Approach for Facility Management Requests Utilizing Association Rule Mining and Large Language Models. *machine learning and knowledge extraction, 6*, 233-258. doi:https://do.org/10.3390/make6010013

[9] Nguyen, T. T., Mullaputi, P. S., Yue, X., Mane, H., Santhos, A., Dennard, E., . . . Nguyen, Q. C. (2026). A decade of discourse: Exploring sentiments and trends around immigration on social media from 2014 to 2024. *Social Science and Medicine*.

[10] Nip, J. Y., & Berthelier, B. (2024). Social media sentiment analysis. In *Encyclopedia* (Vol. 4, pp. 1590–1598). doi:https://doi.org/10.3390/encyclopedia4040104

p-ISSN: 2581-2769; e-ISSN: 2598-9502
DOI: http://dx.doi.org/ 10.32493/JJDP.v9i2.59252

[11] Oller, J. S., & Spadavecchia, C. (2024). Migration and emotions in the media: can socioeconomic indicators predict emotions in images associated with immigrants? *Journal of Computational Social Science*, 963-994.
[12] Pavaloaia, V. D. (2024). Clustering algorithms in sentiment analysis techniques in social media. *International Journal of Advanced Computer Science and Applications, 15*(3), 123–130.
[13] Rasianto, R., & Sutedi, S. (2023). *Association rule mining using FP-growth algorithm for pattern discovery in textual data*.
[14] Regina, E. Sutinah, and N. Agustina, "Clustering Kualitas Kinerja Karyawan Pada Perusahaan Bahan Kimia Menggunakan Algoritma K-Means," J. Media Inform. Budidarma, vol. 5, no. 2, p. 573, 2021, doi: 10.30865/mib.v5i2.2909.
[15] Rogers, R. (2021). Digital methods for social science research. *SAGE Publications*.
[16] Sabna, E., Mustika, B., Fonda, H., Irfan, D., Hang, S., & Pekanbaru, T. (2020). Text Mining Menggunakan Algoritma K-Means Clustering Untuk Memprediksi Keinginan Pasar Terkait Perjalanan Wisata Text Mining Uses K-Means Clustering Algorithm To Predict Market Desires for Tourism Travel. *Journal of Information Technology and Computer Science (INTECOMS), 3*(2), 380–386.
[17] Saputra, A., et al. (2023). *Application of k-means clustering, association rule mining, and classification algorithms in text mining analysis*.
[18] Sugiyono. (2020). *Metode penelitian kuantitatif, kualitatif, dan R&D.* Bandung: Alfabeta.
[19] tim CNN. (2025). *Survei #KaburAjaDulu: Mayoritas Gen Z Ingin Pindah ke Luar Negeri.* CNN Indonesia. Retrieved March 09, 2025, from https://www.cnnindonesia.com/gaya-hidup/20250309155529-277-1206753/survei-kaburajadulu-mayoritas-gen-z-ingin-pindah-ke-luar-negeri#:~:text=Survei%20%23KaburAjaDulu:%20Mayoritas%20Gen%20Z%20Ingin%20Pindah%20ke%20Luar%20Negeri
[20] Vicari, S., & Murru, M. F. (2020). One platform, a thousand worlds: On Twitter irony in the Catalan independence debate. *New Media & Society, 22*(5), 797–818. https://doi.org/10.1177/1461444819879132
[21] World Bank. (2019). *Migration and development brief 31*. Washington, DC: World Bank. Retrieved from Migration and development brief 31.
[22] Zhang, X. (2023). Text classification algorithms exploration on sentiment analysis. *Applied and Computational Engineering, 5*, 99-103.