NAIVE BAYES CHILDFREE

by Sniftyska.Edu

Submission date: 31-Jan-2024 09:48PM (UTC+1100)

Submission ID: 2282802356

File name: NAIVE_BAYES_CHILDFREE.docx (517.89K)

Word count: 5412

Character count: 30692

Abstract— **The difference in societal perspective regarding personal well-being and understanding life choices is genuinely diverse. Lately, there is a prevalent thought where individuals believe that personal well-being can be achieved by choosing to live without children. Most of them prefer to prioritize their careers, education, or other activities that they believe can bring greater happiness and well-being to their lives. This topic has become a frequently discussed subject in almost every region of Indonesia, especially in urban areas. Not only facing negative stigma, the choice to live a life without children in Indonesia also carries positive connotations. Views on child-free in Indonesia are highly diverse, considering the many differences in social environments and each individual's personal experiences. In this research, the Naïve Bayes algorithm is used as a sentiment classifier in the form of textual data collected through Twitter using the Rapid Miner. The data collection period spanned from May 3rd to May 10th, 2023. The research aims to analyze and present data regarding public sentiment towards the child-free phenomenon in Indonesia. The results of the research reveal the presence of 320 positive sentiments and 180 negative sentiments, with the accuracy value of the Naïve Bayes algorithm in conducting sentiment analysis on the child-free phenomenon reached 95.00%.**

Keywords*— *Child-free, Naïve Bayes, Sentiment Analysis, Twitter

I. Introduction

Along with the technological development, particularly on social networks, Indonesia is currently preoccupied with a new "child-free" phenomenon that emerges amidst community life. Child-free is a form of life choice without being blessed with a child. In other words, child-free is a voluntary lifestyle choice to not have children with environmental considerations [1]. The child-free phenomenon, originating from foreign cultures, is slowly becoming present and developing in Indonesia. However, given the diversity of people's perspectives and the numerous differences in people's social life environments, this phenomenon is still widely debated. Some consider this phenomenon suitable to be implemented in Indonesia, and vice versa, considering the cultural context here in Indonesia. Additionally, some individuals within society still need help to filter this phenomenon's presence properly. This is shown by the existence of some people who follow this trend without fully understanding it [2].

The urgency of investigating the child-free phenomenon in Indonesia extends beyond individual choices, but it also encompasses the broader implications on societal structures, cultural norms, and population dynamics. Negative stigma towards childlessness persists, creating challenges for those who choose child-free. This research lies in understanding the individual motivations behind the child-free choice, addressing the societal implications, and fostering a more inclusive understanding of diverse family structures in Indonesia. The importance of research on the "child-free" phenomenon in Indonesia is not only reflected in the individuality of the decision not to have children but also from a broader perspective of its impact on social structures, cultural norms, and the birth rate in the country. Therefore, this clarifies the urgency of a deeper understanding of factors influencing the "child-free" decision, such as career pressures, personal preferences, and the impact of childhood trauma [3].

In the research titled "Being a Child-free Man in Indonesia: Facing Challenges and Social Stigma in Choosing the Freedom Without Children," conducted by Encup Supriatna explained that the child-free phenomenon, particularly from the perspective of men in Indonesia has urgency because it provides a more holistic view of the factors that motivate their decision not to have children, face social stigma, and the strategies they apply to maintain their life choices. This gives valuable insights into a deeper understanding of how social norms and gender expectations influence individuals in making decisions related to family life [4].

One social platform where many people share their comments or express their opinions is Twitter. Tweets written by the public are valuable because they provide feedback on a particular matter. Moreover, these tweets can also be used or serve as a basis for understanding the sentiments or perceptions of the public regarding what is being discussed [5], [6]. This is due to Twitter's expansive user base, exceeding 140 million active users, enabling the daily dissemination of over 400 million tweets or expressions. In addition, Twitter users can engage in discussions on trending topics, facilitating the sharing of insights and gathering feedback from fellow users [7].

The tweet data will undergo processing using text mining methodology to facilitate comprehensive analysis. Text mining, as a discipline, is geared towards the identification and previous disclosure or revelation of heretofore undiscovered yet potentially valuable insights embedded within unstructured or semi-structured textual data [8]. Given the tweets on Twitter, the analysis process naturally requires considerable time. Therefore, several methods can be employed to shorten the time of the sentiment analysis process on the tweet data. One method that can be used to analyze the sentiment of a matter is the Naïve Bayes method or algorithm [9], [10]. The Naïve Bayes algorithm is based on the Bayes principle, which explains that all events have contributions of equal importance or independence concerning selecting a particular class. This algorithm is employed in the text-mining process to visualize societal sentiment [11].

Sentiment analysis is a method of identifying sentiment in the form of text data and how it can be categorized as either positive or negative sentiment. This analytical approach aims to discern and evaluate the emotional tone, attitudes, or opinions expressed within the text, providing insights into the overall subjective context [12], [13]. In other words, sentiment analysis is also referred to as opinion mining. Opinion mining itself is a combination of text mining and natural language processing. The purpose of text mining is to augment textual data originating from specific files and identify the words that represent the content of those files. Consequently, this enables an interconnected analysis between the files [14]. The analysis process will be performed utilizing the Rapid Miner tool, where the Naïve Bayes algorithm will be applied and implemented. In the context of this research, Rapid Miner plays a dual role as both a data analyzer and a data mining engine. It serves as a comprehensive data analysis tool and an engine for the data mining process within the research framework [15]. Within the Rapid Miner platform, several stages are undertaken to align with the objectives of this research. It starts by extracting and aggregating data in the form of opinions from Twitter. After completing the data crawling phase, several stages are implemented to clean the data for optimal analysis. This involves critical stages, such as tokenization, case folding, and stopword removal, to improve the dataset's quality. Following data preprocessing, the following steps involve labeling and classification, an essential component in the analytical pipeline [16], [17]. Following these stages, the analysis is carried out using Naïve Bayes, which allows classification based on the assumption that each predicted attribute has an independent conditional relationship in each class. Consequently, this algorithm is highly effective, yielding robust classification outcomes [18]. This method has proven effective in classifying and performing better than the other methods. Naïve Bayes is better because of its speed and simplicity in classifying text data [19]. Another advantage provided by the Naïve Bayes algorithm is that there is no need for a large amount of training data, so the text classification process to be predicted can be done easily and quickly. To calculate the classification of this method is considered through probability calculations [20].

Lopamudra Dey conducted similar research entitled "Sentiment Analysis of Review Datasets using Naïve Bayes and K-NN Classifier," which discusses evaluating the performance of sentiment classification regarding accuracy and precision value. Both algorithms used to analyze the research topic were Naïve Bayes and K-Nearest Neighbor. The experimental results in this research present that the results with the Naïve Bayes approach are better than the K-NN approach by producing an accuracy value of 80% [21]. Not only that, but other similar research also conducted by Muhammad Andi Ramadhan and Muhammad Iwan Wahyudin entitled "Analisis Sentimen Mengenai Keberhasilan Indonesia di Ajang Thomas Cup 2020 (Studi Kasus Media Sosial Twitter) Menggunakan Metode Naïve Bayes dan Decision Tree". They conducted sentiment analysis on the success of the Indonesian men's badminton team in the 2020 Thomas Cup using Twitter data by comparing two algorithms, which are Naïve Bayes and Decision Tree. According to the research, Naïve Bayes achieved an accuracy rate of 95.161%, while Decision Tree reached 84.677%. This comparison underlines that the Naïve Bayes algorithm in analyzing data is more effective than the Decision Tree [22]. Another similar research that compares Naïve Bayes and another algorithm also conducted by Muhammad Yasir and Robertus Suraji entitled "Perbandingan Metode Klasifikasi Naïve Bayes, Decision Tree, Random Forest terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube". The research conducted by Yasir and his colleague focuses on comparing the performance of Naïve Bayes, Decision Tree, and Random Forest in conducting sentiment analysis on YouTube comments regarding the increase in Hajj pilgrimage costs in 2023. The results of this research indicate that Naïve Bayes achieved accuracy rates of 90%, while Decision Tree achieved 83% and Random Forest achieved 87%. This indicates that Naïve Bayes is more effective in analyzing sentiment than other algorithms [23].

This research diverges from previous research, especially by Muhammad Andi Ramadhan and his colleague, in its approach to sentiment measurement and classifier algorithm as well as the purpose of the research. Unlike the previous research, which measured positive, negative, and neutral sentiment and compared Naïve Bayes with K-NN, Decision Tree, and Random Forest, this research focuses on positive and negative sentiment, utilizing only the Naïve Bayes algorithm as a sentiment analyzer. The key distinction lies in the subject matter and objectives of the research. In addition, this research also incorporates data visualization techniques such as confusion matrix and word cloud, providing a richer visual representation compared to some of the previous research mentioned earlier. As such, this research makes a unique and more comprehensive contribution to understanding the child-free phenomenon in Indonesia.

In this research, the data collected comes from a platform called Twitter, the social media that contains comments or opinions of the public regarding child-free. After that, the data that has been collected will be carried out to a data cleaning process. Subsequently, it will be analyzed using the Naïve Bayes algorithm. This research aims to find out the positive and negative sentiments of Indonesians regarding the child-free phenomenon using tweets uploaded on Twitter. Afterward, the data that has been identified, whether positive or negative, will be analyzed by applying the Naïve Bayes algorithm [20]. This research focuses on public opinion about child-free, which has as many as 500 data on Twitter. Data retrieval and sentiment analysis are carried out using the Rapid Miner tool.

II. METHODOLOGY

2.1 Research Stages

Figure 1 explains the stages of this research, which begin with the data collection (crawling) phase and proceed to the analysis stage of the data, which has gone through the previous stages.

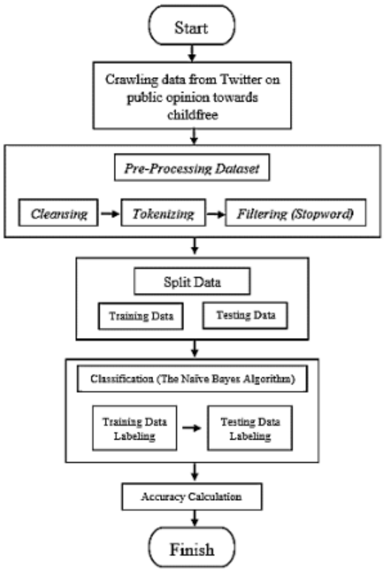


Figure 1. Research Stages

A. Crawling Data from Twitter

This stage marks the initial step, namely, collecting data in the form of opinions from Twitter through Rapid Miner using the Twitter Search operator. The data for this research amounted to 2000 entries, collected from May 3rd to May 10th, 2023. The reason for collecting data during this period is that discussions about child-free in the Twitter community have again become widely discussed.

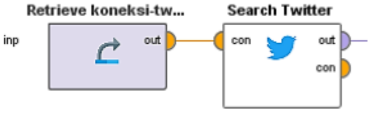
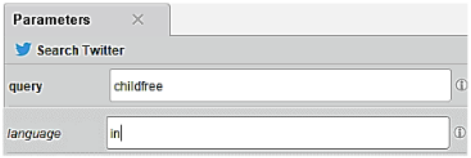


Figure 2. Data Collecting Stage (Crawling Data from Twitter)

In Figure 2, a data collection process through Rapid Miner utilizes two leading operators: Retrieve operator and the Search operator. Since this research aims to gather data from Twitter, the Twitter-connection Retrieve operator employs to connect with Twitter to collect the data that is relevant to this research. On the other hand, the Search Twitter operator is an operator that is utilized to search for tweets on Twitter based on specific keywords, timeframe, and others entered into the parameters of this operator. In this research, the authors collected data using the Search Twitter operator based on the keyword "child-free" entered into the parameter. Furthermore, language regulations were implemented by designating Indonesian as the language parameter, as the objective was to collect tweets in Bahasa Indonesia. This ensured that the collected tweets could be analyzed according to the needs and objectives of the research, as illustrated in Figure 3.


Parameters

Search Twitter

query: childfree

language: in

Figure 3. Parameters in Twitter Search Operator

B. Preprocessing

After the data is collected, a preprocessing stage is carried out, which includes data cleansing, tokenizing, and filtering (stopword removal). This stage is carried out to process the classification or analysis process efficiently [24]. Below is the operator used for performing the data-cleaning process. This operator is the Replace operator, which has a function to cleanse the data from several symbols, numbers, links, and other irrelevant elements. In the data cleaning process of this research, the Replace operator is employed to eliminate terms like "RT", links, mentions, hashtag symbols, and other miscellaneous symbols.

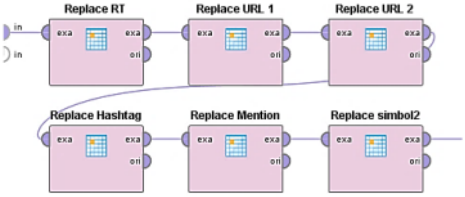


Figure 4. Data Cleansing

The next stage involves tokenizing, case folding, and stopword filtering. The Tokenize operator shown in Figure 5 below is an operator that is utilized as a tool to break or separate text, which may be in the form of paragraphs, sentences, or words that contain symbols into essential words. The Transform Case operator is an operator that is employed to convert all letters in the text into lower case or upper case letters [25]. Meanwhile, the Stopword Filtering operator is used to eliminate words in stopwords. The words in the stopword list are those deemed to lack meaningful contributions to the analysis process. The three operators shown in Figure 5 are included in the operator, namely, Process Document from Data. Process Document from Data is an operator designed to prepare the textual data that will be further analyzed using the chosen algorithm.

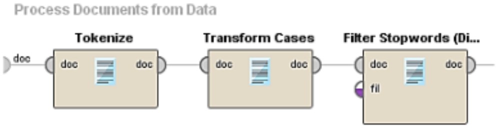


Figure 5. Tokenizing, Case Folding, Stopword

The preprocessing stage involves cleansing, tokenizing, and filtering the data to prepare it for analysis. In addition, removing duplicate entries ensures that the dataset is free of redundant information, resulting in a more focused and efficient analysis. In this case, the dataset, initially 2000 data, has been refined to 500 data ready for further processing. Several examples of data results that have undergone the preprocessing stage are shown in Table 1.

TABLE 1. PREPROCESSING RESULT

Before	After
RT @fatimah_rusalka: @skripssweet Orang tua problematik melahirkan dan membesarkan anak yg problematik juga. Tolongnkalian child-free aja da…	orang tua problematik melahirkan dan membesarkan anak yg problematik juga tolong kalian childfree aja da
itu pilihan pastinya, dan kita ga bisa nilai sepihak kalau itu baik ataupun tidak, pasti di balik itu semua ada alasan tertentu. Jadi kalau emg pasangan milih child-free selama mereka bahagia itu baik tp kalau beda pendapat itu perlu dibicarakan kembali.	itu pilihan pastinya dan kita ga bisa nilai sepihak kalau itu baik ataupun tidak pasti di balik itu semua ada alasan tertentu jadi kalau emg pasangan milih childfree selama mereka bahagia itu baik tp kalau beda pendapat itu perlu dibicarakan kembali

https://t.co/4QY9vduvSO	
tidak masalah, kalau emang child-free karena masih kurangnya finansial itu lebih baik, daripada anak gak terpenuhi gizi dan lain sebagainya https://t.co/PaJyKv1rna	tidak masalah kalau emang childfree karena masih kurangnya finansial itu lebih baik daripada anak gak terpenuhi gizi dan lain sebagainya
@tanyakanrl Aku salah satu yg kelahiran 2000 yg milih child-free bahkan ga berharap sama pernikahan karena rasa trauma masa lalu ???	aku salah satu yg kelahiran 2000 milih childfree bahkan ga berharap sama pernikahan karena rasa trauma masa lalu
@tanyakanrl ngga sih, pasangan yg memutuskan untuk child-free banyak tapi yg masih pgn punya anak jg ga kalah banyak	ngga sih pasangan yg memutuskan untuk childfree banyak tapi masih pgn punya anak jg ga kalah banyak
@risasasasamiya @tanyakanrl Di kota berlaku mungkin. Kalo di desa kemungkinan besar ga bakal child-free. Orang baru nikah beberapa bulan aja udah pada berdoa supaya isi padahal ekonomi masih belum stabil. Terus juga takut di nyinyirin kalo ga segera hamil.	di kota berlaku mungkin kalo di desa kemungkinan besar ga bakal childfree orang baru nikah beberapa bulan aja udah pada berdoa supaya isi padahal ekonomi masih belum stabil terus juga takut di nyinyirin kalo ga segera hamil
tetap pengen punya banyak anak di tengah gempuran gen z child-free ??	tetap pengen punya banyak anak di tengah gempuran gen z childfree
RT @tsubakiee: @tanyakanrl Gak semua anak 2000an milih child-free nder. Kebanyakan org yg milih child-free tu biasanya dari kalangan yg punya lebih banyak akses ke pendidikan dan informasi dan emg cenderung punya lifestyle yg open-minded dan tbh jumlahnya ga sebanyak itu	anak milih childfree nder kebanyakan milih child-free kalangan akses pendidikan informasi cenderung lifestyle open minded

C. Splitting and Labeling Data

After the preprocessing stage, the data is split into two parts: training and testing data. Subsequently, the training data will go through the labeling stage independently, and later, the data will be used as a reference for the analysis of the testing data, which will be conducted directly in the Rapid Miner tool. In this case, the ratio used for the split data stage is 60:40, where 60% is allocated for training data and 40% for testing data. Thus, there are 300 pieces of training data labeled independently by the author and 200 pieces of data that will be tested using the Naïve Bayes algorithm in Rapid Miner. Tweets labeled as positive are those from the society expressing agreement and acceptance of the child-free culture to be implemented in Indonesia and responding positively to the child-free phenomenon. On the other hand, tweets labeled as negative are those from the society expressing disagreement and rejection of the child-free culture to be implemented in Indonesia and responding negatively to the child-free phenomenon. Examples of independently labeled training data categorized into positive and negative sentiments can be seen in Table 2.

TABLE 2. TRAINING DATA LABELING

Text	Label
orang tua problematik melahirkan dan membesarkan anak yg problematik juga tolong kalian childfree aja da	Pos
itu pilihan pastinya dan kita ga bisa nilai sepihak kalau itu baik ataupun tidak pasti di balik itu semua ada alasan tertentu jadi kalau emg pasangan milih childfree selama mereka bahagia itu baik tp kalau beda pendapat itu perlu dibicarakan kembali	Pos
tidak masalah kalau emang childfree karena masih kurangnya finansial itu lebih baik daripada anak gak terpenuhi gizi dan lain sebagainya	Pos
aku salah satu yg kelahiran 2000 milih childfree bahkan ga berharap sama pernikahan karena rasa trauma masa lalu	Pos
ngga sih pasangan yg memutuskan untuk childfree banyak tapi masih pgn punya anak jg ga kalah banyak	Neg
di kota berlaku mungkin kalo di desa kemungkinan besar ga bakal childfree orang baru nikah beberapa bulan aja udah pada berdoa supaya isi padahal ekonomi masih belum stabil terus juga takut di nyinyirin kalo ga segera hamil	Neg
tetap pengen punya banyak anak di tengah gempuran gen z childfree	Neg

The data presented in Table 3 below is one of the partitioned testing dataset samples. Later on, the data will be classified by the Naïve Bayes algorithm to determine the sentiment prediction. The classification result will help calculate probability values and several other values, such as accuracy, recall, precision, and more.

TABLE 3. SAMPLE OF TESTING DATA

Text	Label
anak milih childfree nder kebanyakan milih childfree kalangan akses pendidikan informasi cenderung lifestyle open minded	?

D. Testing Data Classification (Naïve Bayes)

After labeling all the training data independently, the labeled dataset becomes a critical reference point for performing sentiment prediction and testing on the testing data. This process involves applying the Naïve Bayes algorithm within the Rapid Miner platform.

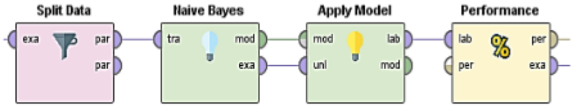


Figure 6. Data Classification (Naïve Bayes)

Figure 6 shows Split Data, Naïve Bayes, Apply Model, and Performance operators. The Split Data operator splits the dataset into: training and testing data. In this research, researchers used a 60:40 ratio in splitting the data, namely 60% for training data and 40% for testing data. In general, dividing the data with a 60:40 ratio can have several advantages, depending on the characteristics of the data and the purpose of the analysis. With a 60:40 ratio, 60% of the data is used to train the model, allowing the model to learn from most of the available data. This can improve the model's ability to capture complex patterns. Additionally, using 40% of the data for testing enables researchers to obtain a more stable evaluation of the model's performance. This ratio can help avoid overfitting and better estimate how the model will perform on new data [26].

Subsequently, the Naïve Bayes classifier algorithm is employed to classify or analyze sentiment. Then, the Apply Model operator is used as a machine learning tool to predict the testing data. The final operator, the Performance operator, is an operator that assesses the performance of the machine learning model. To elaborate, these four operators are the operators that are used in the process of data preparation, model development, model application on test data, and evaluation of model performance in terms of analyzing data. The sentiment predictions for the testing data are obtained through the data classification process, as shown in Table 4.

TABLE 4. NUMBER OF SENTIMENT DATA

Data Type	Positive	Negative	Splitted Data
Training Data	222	78	300
Testing Data	98	102	200
Number of Each Sentiment Type	320	180	500

As shown in Table 4, based on independently labeled training data and data classified by Naïve Bayes, it resulted in 320 data with positive sentiment and 180 data with negative sentiment from the separated training and testing data with a ratio of 60:40, which comprises 300 training data and 200 testing data. From the training data, 222 positive and 78 negative sentiments were generated. Meanwhile, 98 positive and 102 negative sentiments were obtained from the testing data. Thus, it is evident that the total data used in this research amounted to 500 data.

E. Evaluation of Test Results

After completing the testing stage and predicting the sentiment, the next step is calculating the probabilities. Two main probability calculations are at play: prior and posterior probability. It is essential to note that the prior probability's calculation precedes the posterior probability's determination. This sequential process is fundamental as it establishes the foundational probabilities required for subsequent analysis and ensures a comprehensive understanding of the sentiment prediction results.

Calculating prior probabilities is essential for each class in the dataset employed in this research. The strategy is formulated to diminish or alleviate classification bias by taking into account the distribution of classes, spanning from frequently observed to less frequent occurrences. The prior probability calculation generates positive and negative class probabilities derived from the identification process related to the child-free phenomenon. This ensures a balanced consideration of class occurrences and increases the robustness of the subsequent analysis [27]. However, the most crucial aspect is that Naïve Bayes is a classifier algorithm that employs probability calculations. Here are formulas to calculate the prior probability values [14]:

a. Positive Prior Probability

$$\mathrm{P(Positive)} = \frac{\text{Number of Positive Data}}{\text{Total Amount of Data}} \qquad (1)$$

b. Negative Prior Probability

$$\mathrm{P(Negative)} = \frac{\text{Number of Negative Data}}{\text{Total Amount of Data}} \qquad (2)$$

After obtaining the prior probabilities, the subsequent stage involves the calculation of posterior probability. Calculating these posterior probabilities is to ascertain the class for new case identification. The calculation of posterior probabilities in this research is essential for extending the model's capability to accurately classify and handle new cases in the dataset [28]. Here are formulas to calculate the posterior probability values [14]:

a. Positive Posterior Probability

$$\mathrm{Pos.Posterior} = \mathrm{P(X|Pos)} \times \mathrm{P(Positive)} \qquad (3)$$

b. Negative Posterior Probability

$$\mathrm{Neg.Posterior} = \mathrm{P(X|Neg)} \times \mathrm{P(Negative)} \qquad (4)$$

Besides the probability values, accuracy, precision, and recall values were also analyzed and calculated. These values can only be calculated by first calculating the confusion matrix in Table 5 and Figure 7.

TABLE 5. CONFUSION MATRIX

Prediction	True	
	Positive	**Negative**
Positive	89 (TP)	1 (FN)
Negative	9 (FP)	101 (TN)

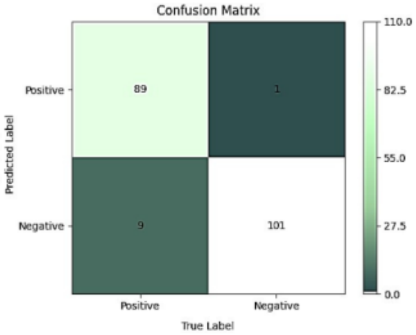


Figure 7. Confusion Matrix Visualization

The confusion matrix depicted in Table 5 and Figure 7 above includes an account of the classification of data amounts that are accurately predicted as positive sentiment or True Positive (TP), erroneously predicted as positive sentiment or False Positive (FP), accurately predicted as negative sentiment or True Negative (TN), and erroneously predicted as negative sentiment or False Negative (FN) derived from testing data. Subsequently, the confusion matrix will be utilized to compute values such as accuracy, recall or True Positive Rate (TPR), precision or Positive Predictive Value (PPV), True Negative Rate (TNR), and Negative Predictive Value (NPV).

The accuracy value signifies the percentage of testing data whose class has been correctly identified or classified by the system based on their original class. The formula to calculate the accuracy value is as follows [29]:

$$\textbf{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \tag{5}$$

The recall value, also known as True Positive Rate (TPR), has another name: sensitivity value, which represents the percentage of the system's success in classifying positive class as positive class. The formula for calculating the recall value (TPR) is given in equation (6) below [29]:

$$\textbf{Recall (True Positive Rate)} = \frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

The precision value, also known as Positive Predictive Value (PPV), represents the percentage of data predicted as a positive class by the classification algorithm, which is the actual positive data of all those predicted as positive class. The formula for calculating the precision value (PPV) is given in equation (7) below [29]:

$$\textbf{Precision (Positive Predictive Value)} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

The True Negative Rate (TNR) value, which has another name, namely, the specificity value, represents the percentage of the system's success in correctly classifying the negative class as a negative class. The True Negative Rate (TNR) value calculation formula is given in equation (8) below [30]:

$$\textbf{True Negative Rate} = \frac{\text{TN}}{\text{FP+TP}} \tag{8}$$

The NPV or Negative Predictive Value represents the percentage of data predicted as a negative class by the classification algorithm that is correctly negative data out of all classes predicted as negative. The formula for calculating the Negative Predictive Value (NPV) is given in equation (9) below [30]:

$$\textbf{Negative Predictive Value} = \frac{\text{TN}}{\text{FN+TN}} \tag{9}$$

III. Result and Analysis

This section presents the results of the research and various tests. The calculation of probabilities follows the sentiment analysis of the data using the Naïve Bayes algorithm. The probability calculations employed in this research include prior probability and posterior probability. After calculating the prior and posterior probability values, the accuracy, recall (TPR), precision (PPV), TNR, and NPV values are also calculated using the confusion matrix as a reference in performing the five calculations after the probability calculations.

It is necessary to identify the most frequently occurring words from the data collected to calculated both probability values: prior and posterior. Below is the word cloud visualization of the most commonly appearing words in the data collected for this research:

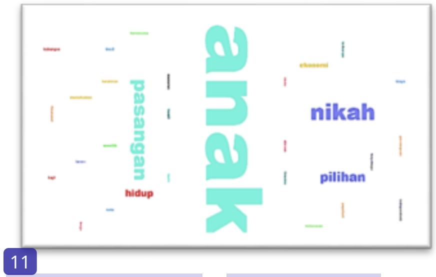


Figure 8. Word Cloud Visualization of the Most Commonly Appearing Words

From the word cloud presented in Figure 8 above, three frequently occurring words will be selected as examples and references for calculating prior and posterior probabilities. These selected words include "anak," which appears 125 times, "nikah," which appears 37 times, and "pasangan," which appears 30 times. The sentiment examples that were chosen for the calculation of prior and posterior probabilities are shown in Table 6 below.

TABLE 6. DATA CLASSIFICATION

	Anak	Nikah	Pasangan	Label
P2	1	0	0	Pos
P62	0	0	1	Pos
P69	1	0	0	Pos
P99	0	1	0	Pos
N36	1	0	1	Neg
N48	0	1	0	Neg
N67	1	0	0	Neg
U169	1	0	0	?

Here is the calculation of the prior and posterior probabilities of some data that has been classified:

1. Prior Probability

$$\mathbf{P(Positive)} = \frac{\text{Number of Positive Data}}{\text{Total Amount of Data}} = \frac{4}{7} = 0{,}57$$

$$\mathbf{P(Negative)} = \frac{\text{Number of Negative Data}}{\text{Total Amount of Data}} = \frac{3}{7} = 0{,}42$$

From those values, the value of the positive class and negative class can be calculated as shown below [14]:

Positive Class (P(X-169|Pos))
= P(anak=2|Pos) × P(nikah=1|Pos) × P(pasangan=1|Pos)
= 0,5 × 0,25 × 0,25
= 0,03125

Negative Class (P(X-169|Neg))
= P(anak=2|Neg) × P(nikah=1|Neg) × P(pasangan=1|Neg)
= 0,66 × 0,33 × 0,33

= 0,07187

2. Posterior Probability

Pos. Posterior = (P(X-169|Pos) × P(Positive)
= 0,03125×0,57
= 0,0178125

Neg. Posterior = (P(X-169|Neg) × P(Negative)
= 0,07187×0,42
= 0,0301854

After calculating these probability values, it can be classified that the 169th testing data falls into the negative sentiment category. The resulting negative posterior probability value is greater than the positive posterior probability.

TABLE 7. CLASSIFIED TESTING DATA

Text	Label
anak milih child-free nder kebanyakan milih child-free kalangan akses pendidikan informasi cenderung lifestyle open minded	Negative

From Table 4, the confusion matrix, calculations are made for the accuracy, recall (TPR), precision (PPV), TNR, and NPV values. Table 7 shows an example of classified testing data classified as negative sentiment based on calculating both probability values, namely, prior probability and posterior probability. The accuracy value calculation is carried out to assess the extent to which the Naïve Bayes classification algorithm can correctly predict the entire testing dataset. Furthermore, the recall value or True Positive Rate (TPR) is calculated to assess how far the Naïve Bayes classifier algorithm can detect all positive sentiments correctly. Subsequently, the precision value or Positive Predictive Value (PPV) is calculated to measure how far the Naïve Bayes classifier algorithm can provide positive predictions accurately. Following that, the True Negative Rate (TNR) value calculation is used to assess how far the Naïve Bayes algorithm can accurately detect all negative sentiments. Lastly, the Negative Predictive Value (NPV) is calculated to evaluate how accurately the Naïve Bayes classification algorithm can provide negative predictions. Below is the calculation for accuracy, recall, precision, TNR, and NPV values:

a. Accuracy

$$\text{Accuracy} = \frac{89+101}{89+9+1+101}$$

$$\text{Accuracy} = 0{,}95 \text{ or } 95\%$$

b. Recall (True Positive Rate)

$$\text{Recall} = \frac{89}{89+1}$$

$$\text{Recall} = 0{,}9889 \text{ or } 98{,}89\%$$

c. Precision (Positive Predictive Value)

$$\text{PPV} = \frac{89}{89+9}$$

$$\text{PPV} = 0{,}9082 \text{ or } 90{,}82\%$$

d. TNR (True Negative Rate)

$$\text{TNR} = \frac{101}{9+101}$$

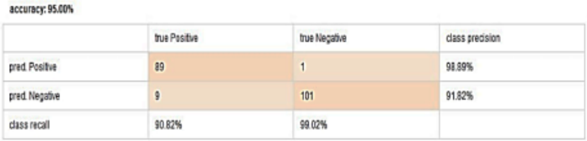
$$\text{TNR} = 0{,}9182 \text{ or } 91{,}82\%$$

e. NPV (Negative Predictive Value)

$$\text{NPV} = \frac{101}{101+1}$$

$$\mathrm{NPV} = 0{,}9902 \text{ or } 99{,}02\%$$

From these calculations, it can be explained that the Naïve Bayes algorithm effectively analyzes the sentiment. This is evidenced by calculating the accuracy value, which reaches 95.00%. This means that the algorithm or classification model has a high level of accuracy in predicting positive and negative sentiments correctly and accurately. Furthermore, the results of calculating the recall (TPR) and TNR values are 98.89% and 91.82%, respectively. This implies that the recall (TPR) and TNR values clarify that the classification algorithm or model can correctly detect positive and negative sentiments. The subsequent calculation is the calculation of the precision (PPV) and NPV values, yielding 90.82% and 99.02%, respectively. From these calculations, it can be seen that the classification algorithm or model indicates that each predicted positive and negative sentiment is correct or accurate.



accuracy: 95.00%

	true Positive	true Negative	class precision
pred. Positive	89	1	98.89%
pred. Negative	9	101	91.82%
class recall	90.82%	99.02%	

Figure 9. Calculation Results on Rapid Miner

The evaluation results using the Naïve Bayes classification algorithm with 95.00% accuracy, 98.89% recall, 90.82% precision, 91.82% True Negative Rate (TNR), and 99.02% Negative Predictive Value (NPV) indicate the effectiveness of the algorithm in predicting positive and negative sentiments related to the child-free phenomenon in Indonesia. The high level of accuracy, both in detecting positive and negative sentiments, provides empirical support for the urgency of this research. These results imply a deep understanding of individual child-free life choices and reflect broad societal support and cultural norms in Indonesia. Therefore, this study provides a strong foundation for understanding the impact of the child-free phenomenon on social structures, cultural norms, and population dynamics in Indonesia.

IV. Conclusion

From the results of this research, it can be concluded that the application of the Naïve Bayes classification algorithm in analyzing the sentiment of Indonesian public opinion regarding the child-free phenomenon, based on data collected from Twitter, as much as 500 data that has gone through the cleaning stage, has proven effective with an accuracy value of 95.00%. This shows that the Naïve Bayes algorithm can predict sentiment or public opinion about the child-free phenomenon precisely and accurately. Furthermore, the analysis results also show that public opinion tends to be more positive than negative towards the child-free phenomenon in Indonesia, with more positive sentiments than negative sentiments, with 320 and 180 negative sentiments, respectively. This illustrates the tendency of Indonesian people's views towards this phenomenon. This is because child-free can be understood as a reflection of the Indonesian people's views on child-free life, most of which tend to support and accept the concept of child-free. Thus, the child-free phenomenon in Indonesia is not only an individual life choice but also reflects broader social support and understanding.

As a direction for future research, the analysis may be extended towards temporal dimensions, demographic variations, multilingual studies, and the inclusion of qualitative methods and ethical considerations to deepen understanding and maintain ethical aspects in sentiment analysis. Thus, future research can provide more comprehensive insights into the influence of the child-free phenomenon on social structures, cultural norms, and population dynamics.

NAIVE BAYES CHILDFREE

ORIGINALITY REPORT

6%	3%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	etheses.uinsgd.ac.id Internet Source	<1%
2	www.scirp.org Internet Source	<1%
3	repositorio.upct.es Internet Source	<1%
4	www.scilit.net Internet Source	<1%
5	Enrico Corradini. "The Dark Threads That Weave the Web of Shame: A Network Science-Inspired Analysis of Body Shaming on Reddit", Information, 2023 Publication	<1%
6	Adian Fatchur Rochim, Khoirunisa Widyaningrum, Dania Eridani. "Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment", 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2021 Publication	<1%

7	Arcadius Benawa. "The Significance Influence of Trust and Organizational Commitment on The SPIRIT Characters Building for The Students", 2023 2nd Asia-Pacific Computer Technologies Conference (APCT), 2023 Publication	<1 %
8	Rahmat Syahputra, Gomal Juni Yanris, Deci Irmayani. "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter", Sinkron, 2022 Publication	<1 %
9	ebin.pub Internet Source	<1 %
10	www.mdpi.com Internet Source	<1 %
11	Ogle, David Alex. "Personalized Learning: A Meta-Analysis", The Southern Baptist Theological Seminary, 2021 Publication	<1 %
12	1library.net Internet Source	<1 %
13	publications.eai.eu Internet Source	<1 %
14	link.springer.com Internet Source	<1 %

15	D T Pham, S S Dimov, C D Nguyen. "A two-phase K-means algorithm for large datasets", Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005 Publication	<1 %
16	ojs.cahayamandalika.com Internet Source	<1 %
17	www.jait.us Internet Source	<1 %
18	Ali Mustopa, Hermanto, Anna, Eri Bayu Pratama, Ade Hendini, Deni Risdiansyah. "Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naive Bayes Algorithm Based on Particle Swarm Optimization", 2020 Fifth International Conference on Informatics and Computing (ICIC), 2020 Publication	<1 %
19	Khadijah, Retno Kusumaningrum, Rismiyati, Auliya Mujadidurrahman. "An Efficient Masked Face Classifier Using EfficientNet", 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), 2021 Publication	<1 %

20	Zhongdi Wu, Stuart Stothoff, Osvaldo Pensado, Jennifer Alford, Eric C. Larson. "Exploring Convolutional Neural Networks for Predicting Sentinel-C Backscatter between Image Acquisitions", IEEE Transactions on Geoscience and Remote Sensing, 2023 Publication	<1 %
21	www.ijraset.com Internet Source	<1 %
22	Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari. "Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier", International Journal of Information Engineering and Electronic Business, 2016 Publication	<1 %
23	journal.lembagakita.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	Off		

NAIVE BAYES CHILDFREE