Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

Sentiment Analysis of Reviews of Tourist Attractions in the Lake Toba Area Using the Naïve Bayes Method

Yuke Wiranti[1)*], Yusuf Ramadhan Nasution[2)]
[1)*2)]Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[1)*]yukewiranti37@gmail.com, [2)]ramadhannst@uinsu.ac.id

ABSTRACT

Lake Toba is one of the tourism destinations in Indonesia which is the main destination for domestic and foreign tourists. However, the natural beauty of Lake Toba is not enough to develop quality tourist destinations, so analysis needs to be carried out in order to develop tourist destinations that suit tourist needs with the aim of improving the economy from tourism, especially at Lake Toba. One aspect that must be analyzed is comments from tourists who have visited Lake Toba via various platforms. This is very influential for potential future tourists to have a reference for Lake Toba tourism. The analysis process can be carried out by analyzing comments using the Naive Bayes method so that managers of the Lake Toba tourist destination can improve tourist attractions and provide tourist satisfaction and develop various tourism innovations to meet various tourist needs in a sustainable manner. The results of the sentiment analysis of Lake Toba tourist reviews using Naive Bayes detected 31 positive labels, 378 neutral labels and 7 negative labels with an accuracy result of 77.49% from 1260 data, where training data was 1008 and test data was 252 data.

Keywords: Sentiment Analysis; Review; TF-IDF; Naïve Bayes

1. INTRODUCTION

Indonesia is a country that has a comparative advantage in beautiful natural landscapes, one of which is Lake Toba. Lake Toba is one of the tourism destinations which is the main destination for domestic and foreign tourists in Indonesia. Currently, Lake Toba has been designated as a Super Priority Destination by the Government as stated in the Decree of President Joko Widodo at the limited cabinet meeting on 15 July 2019. (Saragih et al., 2021). However, the natural beauty of Lake Toba is not enough to develop quality tourist destinations, so analysis needs to be carried out in order to develop tourist destinations that suit tourist needs with the aim of improving the economy from tourism, especially at Lake Toba.

One aspect that must be analyzed are various comments from tourists who have carried out tourism activities at the Lake Toba tourist destination, such as comments on social media. Through social media, many tourists share their experiences about tourist attractions, one of which is Lake Toba, through various platforms such as the TripAdvisor website, social media or by filling in Google From, one of which is a post from @kesamosir_aja on July 24 2024 which states ”Danau Toba terancam kehilangan status Global Geopark dari UNESCO” and commented on the account @HARY93 which filled in the comments ”pinggiran Danau Toba bauk amis, mau berenangpun jijai”.

This is very influential for potential future tourists because these comments can be used as a reference for tourist attractions, especially Lake Toba. Therefore, it is necessary for managers of Lake Toba tourist attractions to see the reviews made by each tourist to be able to know the overall purpose of the reviews that have been made. Of course, the number of opinions published on social media is too large to be processed manually, so a system is needed that is able to automatically analyze the number of reviews made by tourists about tourist attractions on Lake Toba..

The process can be carried out by analyzing the ratings or scores of sentiment reviews on social media using the Naive Bayes method so that Lake Toba tourist destination managers can improve tourist attractions and be able to provide satisfaction to tourists as well as develop various tourism innovations to meet various tourist needs in a sustainable manner. If the destination is able to provide performance that can meet tourist needs in a sustainable manner, it will have a good impact on the development of tourist destinations in the Lake Toba area.

Sentiment analysis is the process of extracting information obtained from various data sources such as the internet

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

and various social media platforms that form a person's view of an issue (Maulani et al., 2019). The help of sentiment analysis is able to provide information that was initially unstructured into more structured data (Sari & Wibowo, 2019). Sentiment analysis is the process of using text analytics to obtain various data sources from the internet and various social media platforms. The aim is to obtain opinions from users on the platform (Evita Fitri, Yuri Yuliani, Susy Rosyida, 2020). Sentiment analysis can be processed using the Naive Bayes method.

2. LITERATURE REVIEW

Research on sentiment analysis using Naive Bayes has been carried out several times as carried out by (Sidabutar et al., 2023) has an accuracy result of 74.57% on tweet data from Twitter in the Lake Toba area. Then research conducted by (Rahel Lina Simanjuntak et al., 2023) has an accuracy result of 92% on review data on E-Commerce applications with a data ratio of 80:20. Meanwhile, research conducted by (Karyati, 2020) Regarding sentiment analysis on Twitter users' opinions about Gunadarma University using Naive Bayes, it has an accuracy of 86.42% with an error of 13.58%.

3. METHOD

The research method used to produce accurate data from sentiment analysis and test the effectiveness of the method in conducting sentiment analysis on reviews of tourist attractions in the Lake Toba area is Naive Bayes. The steps that can be taken to carry out research and design a simple system are as follows:

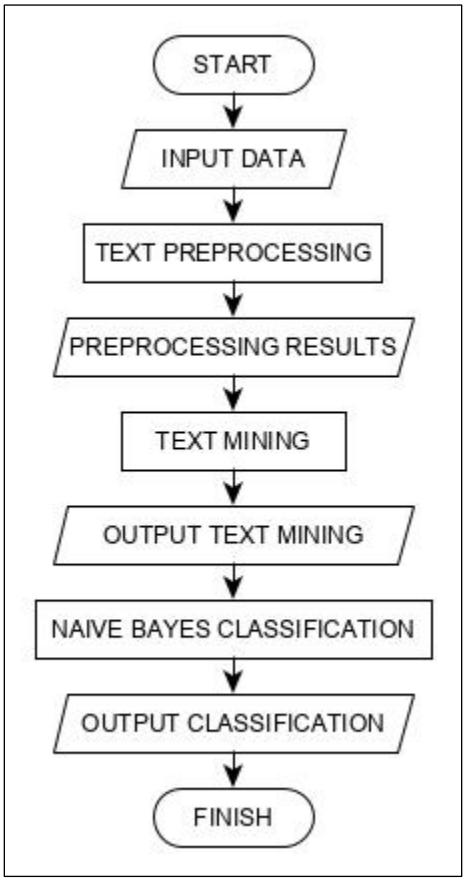

Fig. 1 Classification system flowchart using Naive Bayes

Based on Figure 1, it can be seen that the process that must be carried out after inputting data on reviews of Lake Toba tourist attractions is text preprocessing. Text Preprocessing is a stage in cleaning the dataset before proceeding to the classification stage in the Naive Bayes method (Deolika et al., 2019). Text preprocessing is the process of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

preparing Lake Toba tourist attractions review data that will be used in the knowledge discovery operation of the Text mining system . The actions taken at this stage are toLowerCase, which converts all letter characters into lowercase letters and Tokenizing. Broadly speaking, tokenisation is the stage of breaking down a set of characters in a text into word units. The set of characters can be whitespace characters, such as enter, tabulation, space. But for single quote characters ('), period (.), semicolon (;), colon (:) or others, can also have quite a lot of roles as word separators (Amrullah et al., 2020). The text preprocessing stage in processing text data in the form of reviews of Lake Toba tourist attractions, namely: case folding, tokenizing, filtering and stemming (Fitrianah et al., 2021).

After text preprocessing the data, it will proceed to the process of classifying data on reviews of Lake Toba tourist attractions using the Naive Bayes method. The Naive Bayes method is a classification method that is able to predict future opportunities based on previous experience, known as the Bayes Theorem (Herdhianto, 2020). Naive Bayes classification is a classification that requires a small amount of training data to determine the parameter estimates needed in the classification process (Toy et al., 2021)(Nasution & KHAIRUNA, 2017)The equation in Naive Bayes is as follows:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

Where C_i is hypothesis that data X is a specific category, X is the unknown data class, $P(C_i|X)$ is the probability of hypothesis C_i based on condition X (posteriori probability), $P(X|C_i)$ is the probability of X based on condition in hypothesis C_i, $P(_{Ci})$ is the probability of hypothesis C_i (prior probability) and P(X) expresses the probability of data X.

$$P(X|C) = P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n|C) \quad (2)$$

In practice, where P(X) is constant for all classes, we can ignore the posterior probability formula and compare the relative posterior probabilities for each class to perform classification .

After text preprocessing of the data has been carried out, it will proceed to the data classification process for reviews of Lake Toba tourist attractions using the Naive Bayes method. Prediction results from review data of Lake Toba tourist attractions will be tested for accuracy using a confusion matrix with the following equation (Nitamia & Februariyanti, 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Dimana:
TP = True Positive
TN = True Negative
FN = False Negative
FP = False Positive

4. RESULT

ANALISIS DATA

In conducting sentiment analysis, the author collected review data from tourist attractions in the Lake Toba area. This involved pulling data from travel website platform TripAdvisor, social media TikTok and Google From. After obtaining the required data, the next step is to save the data in Excel with the ".xlsx" format. The total data of Lake Toba tourist reviews is 1260 data consisting of 100 positive reviews, 1135 neutral reviews and 25 negative reviews.

REPRESENTASI DATA

At the Data Representation stage, the stages of explaining the process in the research carried out will be carried out starting from the data collection stage to the stage of applying the Naive Bayes method to produce predictions for sentiment analysis of tourist attractions in the Lake Toba area. The stages are as follows:

a. Labelisasi

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

Labeling the dataset has the function of determining the class (label) for each sentiment. At this stage the sentiment obtained certainly has a positive, negative or neutral value to be given to each sentiment. The following dataset has been labeled as follows:

Table 1
Sentiments that have been labeled

Train Sentiment	Class
Keindahan alam yg mempesona	Positif
Tempatnya sangat bagus dan nyaman	Positif
Pemandangannya indah apalagi kalau cuacanya cerah	Positif
Jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan	Negatif
Biasa aja, tapi hampir semua fasilitas di sana mahal banget	Negatif

b. Text Preprocessing

Text Preprocessing is a stage in cleaning the dataset before proceeding to the classification stage using the Naive Bayes method which is determined first. The preprocessing stage will go through several stages as follows:

Case Folding

Table 2
Stages of case folding

Initial Sentiment	Removal of links, hashtags and emojis
Data Training	
keindahan alam yg mempesona	keindahan alam yg mempesona
tempatnya sangat bagus dan nyaman	tempatnya sangat bagus dan nyaman
pemandangannya indah apalagi kalau cuacanya cerah	pemandangannya indah apalagi kalau cuacanya cerah
jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan	jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan
biasa aja, tapi hampir semua fasilitas di sana mahal banget	biasa aja, tapi hampir semua fasilitas di sana mahal banget
Data Testing	
karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa	karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa
karena wisata danau toba itu merupakan surga dunia	karena wisata danau toba itu merupakan surga dunia

Cleaning

Table 3
Cleaning Stages

Initial Sentiment	Removal of links, hashtags and emojis
Data Training	
keindahan alam yg mempesona	keindahan alam yg mempesona
tempatnya sangat bagus dan nyaman	tempatnya sangat bagus dan nyaman
pemandangannya indah apalagi kalau cuacanya cerah	pemandangannya indah apalagi kalau cuacanya cerah
jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan	jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan
biasa aja, tapi hampir semua fasilitas di sana mahal banget	biasa aja, tapi hampir semua fasilitas di sana mahal banget
Data Testing	
karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa	karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

karena wisata danau toba itu merupakan surga dunia	karena wisata danau toba itu merupakan surga dunia

Number Removal

Table 4
Number Removal

Initial Sentiment	Number Removal
Data Training	
keindahan alam yg mempesona	keindahan alam yg mempesona
tempatnya sangat bagus dan nyaman	tempatnya sangat bagus dan nyaman
pemandangannya indah apalagi kalau cuacanya cerah	pemandangannya indah apalagi kalau cuacanya cerah
jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan	jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan
biasa aja, tapi hampir semua fasilitas di sana mahal banget	biasa aja, tapi hampir semua fasilitas di sana mahal banget
Data Testing	
karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa	karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa
karena wisata danau toba itu merupakan surga dunia	karena wisata danau toba itu merupakan surga dunia

Punctuation Removal

Table 5
Punctuation Removal

Initial Sentiment	Punctuation Removal
Data Training	
keindahan alam yg mempesona	keindahan alam yg mempesona
tempatnya sangat bagus dan nyaman	tempatnya sangat bagus dan nyaman
pemandangannya indah apalagi kalau cuacanya cerah	pemandangannya indah apalagi kalau cuacanya cerah
jelek banget, view nya gitu gitu aja, engga bagus sama sekali gak ada perubahan	jelek banget view nya gitu gitu aja engga bagus sama sekali gak ada perubahan
biasa aja, tapi hampir semua fasilitas di sana mahal banget	biasa aja tapi hampir semua fasilitas di sana mahal banget
Data Testing	
karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa	karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa
karena wisata danau toba itu merupakan surga dunia	karena wisata danau toba itu merupakan surga dunia

Tokenisasi

Table 6
Tokenization Stages

Initial Sentiment	Tokenisasi
Data Training	
keindahan alam yg mempesona	['keindahan', 'alam', 'yg', 'mempesona']
tempatnya sangat bagus dan nyaman	['tempatnya', 'sangat', 'bagus', 'dan', 'nyaman']
pemandangannya indah apalagi kalau cuacanya cerah	['pemandangannya', 'indah', 'apalagi', 'kalau', 'cuacanya', 'cerah']
jelek banget view nya gitu gitu aja engga bagus sama sekali gak ada perubahan	['jelek', 'banget', 'view', 'nya', 'gitu', 'gitu', 'aja', 'engga', 'bagus', 'sama', 'sekali', 'gak', 'ada', 'perubahan']

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

biasa aja tapi hampir semua fasilitas di sana mahal banget	['biasa', 'aja', 'tapi', 'hampir', 'semua', 'fasilitas', 'di', 'sana', 'mahal', 'banget']
Data Testing	
karena danau toba merupakan pesona alami yang menawarkan keindahan yang luar biasa	['karena', 'danau', 'toba', 'merupakan', 'pesona', 'alami', 'yang', 'menawarkan', 'keindahan', 'yang', 'luar', 'biasa']
karena wisata danau toba itu merupakan surga dunia	['karena', 'wisata', 'danau', 'toba', 'itu', 'merupakan', 'surga', 'dunia']

Stopword

Table 7
Stopword Stages

Initial Sentiment	Penghapusan stopword
Data Training	
['keindahan', 'alam', 'yang', 'mempesona']	['keindahan', 'alam', 'mempesona']
['tempatnya', 'sangat', 'bagus', 'dan', 'nyaman']	['tempatnya', 'bagus', 'nyaman']
['pemandangannya', 'indah', 'apalagi', 'kalau', 'cuacanya', 'cerah']	['pemandangannya', 'indah', 'cuacanya', 'cerah']
['jelek', 'banget', 'view', 'nya', 'begitu', 'begitu', 'saja', 'enggak', 'bagus', 'sama', 'sekali', 'enggak', 'ada', 'perubahan']	['jelek', 'view', 'bagus', 'perubahan']
['biasa', 'saja', 'tapi', 'hampir', 'semua', 'fasilitas', 'di', 'sana', 'mahal', 'banget']	['fasilitas', 'mahal']
Data Testing	
['karena', 'danau', 'toba', 'merupakan', 'pesona', 'alami', 'yang', 'menawarkan', 'keindahan', 'yang', 'luar', 'biasa']	['danau', 'toba', 'pesona', 'alami', 'menawarkan', 'keindahan']
['karena', 'wisata', 'danau', 'toba', 'itu', 'merupakan', 'surga', 'dunia']	['wisata', 'danau', 'toba', 'surga', 'dunia']

Stemming

Table 8
Stemming Stages

Initial Sentiment	Proses stemming
Data Training	
['keindahan', 'alam', 'mempesona']	['indah', 'alam', 'pesona']
['tempatnya', 'bagus', 'nyaman']	['tempat', 'bagus', 'nyaman']
['pemandangannya', 'indah', 'cuacanya', 'cerah']	['pandang', 'indah', 'cuaca', 'cerah']
['jelek', 'view', 'bagus', 'perubahan']	['jelek', 'view', 'bagus', 'ubah']
['fasilitas', 'mahal']	['fasilitas', 'mahal']
Data Testing	
['danau', 'toba', 'pesona', 'alami', 'menawarkan', 'keindahan']	['danau', 'toba', 'pesona', 'alami', 'tawar', 'indah']
['wisata', 'danau', 'toba', 'surga', 'dunia']	['wisata', 'danau', 'toba', 'surga', 'dunia']

c. TF-IDF Weighting

After carrying out the text preprocessing stage, the next step is to carry out the weighting stage using TF-IDF in

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

the calculation process. The following are the stages of weighting several words that have been carried out:

TF-DF Weighting

Table 9
Weighting TF-DF Values from Training Data

Term	TF					DF
	D1	D2	D3	D4	D5	
Indah	1	0	1	0	0	2
Alam	1	0	0	0	0	1
Pesona	1	0	0	0	0	1
Tempat	0	1	0	0	0	1
Bagus	0	1	0	1	0	2
Nyaman	0	1	0	0	0	1
Pandang	0	0	1	0	0	1
Cuaca	0	0	1	0	0	1
Cerah	0	0	1	0	0	1
Jelek	0	0	0	1	0	1
View	0	0	0	1	0	1
Ubah	0	0	0	1	0	1
Fasilitas	0	0	0	0	1	1
Mahal	0	0	0	0	1	1

After the TF (term frequency) value is obtained, the next step is to find the value of the IDF. Below is an equation to determine the IDF value of each word.

$$IDF = \ln\left(\frac{D+1}{df+1}\right) + 1 \quad (4)$$

DF-IDF Weighting

Table 10
DF-IDF Values from Training Data

Term	DF	IDF
Indah	2	1.693
Alam	1	2.098
Pesona	1	2.098
Tempat	1	2.098
Bagus	2	1.693
Nyaman	1	2.098
Pandang	1	2.098
Cuaca	1	2.098
Cerah	1	2.098
Jelek	1	2.098
View	1	2.098
Ubah	1	2.098
Fasilitas	1	2.098
Mahal	1	2.098

After the TF and IDF values are obtained, the TF-IDF value can be calculated. To find the TF-IDF value, use the

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

equation below.

$W = TF\ x\ IDF$ (5)

TF-IDF Weighting

Table 11
TF-IDF value from training data

No.	*Term*	TF-IDF				
		D1	**D2**	**D3**	**D4**	**D5**
1	Indah	1.693	0	1.693	0	0
2	Alam	2.098	0	0	0	0
3	Pesona	2.098	0	0	0	0
4	Tempat	0	2.098	0	0	0
5	Bagus	0	1.693	0	1.693	0
6	Nyaman	0	2.098	0	0	0
7	Pandang	0	0	2.098	0	0
8	Cuaca	0	0	2.098	0	0
9	Cerah	0	0	2.098	0	0
10	Jelek	0	0	0	2.098	0
11	View	0	0	0	2.098	0
12	Ubah	0	0	0	2.098	0
13	Fasilitas	0	0	0	0	2.098
14	Mahal	0	0	0	0	2.098

Next, the TF-IDF value is normalized to equalize the interval of each data. The equation used to normalize the data is as follows..

$$TF_{norm}(t,d) = \frac{TF(t,d)}{\sqrt{\sum_i (TF(t,d))^{\ 2}}} \quad (6)$$

The following are the results of the data normalization calculations carried out.

Table 12
Data Normalization

No.	Term	D1	D2	D3	D4	D5
1	Indah	0.2112	0	0.2112	0	0
2	Alam	0.2617	0	0	0	0
3	Pesona	0.2617	0	0	0	0
4	Wisata	0	0.2617	0	0	0
5	Bagus	0	0.2112	0	0.2112	0
6	Nyaman	0	0.2617	0	0	0
7	Pandang	0	0	0.2617	0	0
8	Cuaca	0	0	0.2617	0	0
9	Cerah	0	0	0.2617	0	0
10	Jelek	0	0	0	0.2617	0
11	View	0	0	0	0.2617	0
12	Ubah	0	0	0	0.2617	0
13	Fasilitas	0	0	0	0	0.2617
14	Mahal	0	0	0	0	0.2617

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

d. Naive Bayes Classification

Table 12
Test sentiment

Test Sentiment
['danau', 'toba', 'pesona', 'alami', 'tawar', 'indah']
['wisata', 'danau', 'toba', 'surga', 'dunia']

The test sentiment will be carried out by searching for probability values using equation 1, so that the probability of each class is as follows.

P (Positif | Sentimen) $= \frac{3}{5} = 0.4$
P (Negatif | Sentimen) $= \frac{2}{5} = 0.4$
P (Netral | Sentimen) $= \frac{1}{5} = 0.2$

The next step is to find the conditional probability value using equation 2 as follows.

$$P\ (Term \mid Class) = \frac{Total\ TF{-}IDF\ Term\ Weight\ in\ Class{+}1}{Total\ TF{-}IDF\ Weight} \quad (7)$$

So the conditional probability (term) in each class for the test sentiment is as follows.

For Positive Class

P (danau | Positif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (toba | Positif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (pesona | Positif) $= \frac{0.262+1}{3.985} = \frac{1.262}{3.985} = 0.317$
P (alami | Positif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (tawar | Positif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (indah | Positif) $= \frac{0.211+1}{3.985} = \frac{1.211}{3.985} = 0.303$

For Negative Class

P (danau | Negatif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (toba | Negatif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (pesona | Negatif) $= \frac{0+1}{3.985} = \frac{1.262}{3.985} = 0.251$
P (alami | Negatif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (tawar | Negatif) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (indah | Negatif) $= \frac{0+1}{3.985} = \frac{1.422}{3.985} = 0.251$

For Neutral Class

P (danau | Netral) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (toba | Netral) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (pesona | Netral) $= \frac{0+1}{3.985} = \frac{1.262}{3.985} = 0.251$
P (alami | Netral) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$
P (tawar | Netral) $= \frac{0+1}{3.985} = \frac{1}{3.985} = 0.251$

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

P (indah | Netral) $= \frac{0.211 + 1}{3.985} = \frac{1.211}{3.985} = 0.303$

Then look for the posterior probability value using equation 2, so that the results for each class are as follows.

P (Sentimen Uji 1| Positif) = 0.251 * 0.251 * 0.317 * 0.251 * 0.251 * 0.303 * 0.4
= 0.000152

P (Sentimen Uji 1 | Negatif) = 0.251 * 0.251 * 0.251 * 0.251 * 0.251 * 0.251 * 0.4
= 0.0001000

P (Sentimen Uji 1 | Netral) = 0.251 * 0.251 * 0.251 * 0.251 * 0.251 * 0.303 * 0.2
= 0.0000603

Based on the example calculations carried out, the highest calculated value from all classes and the highest value from the test data on Positive sentiment were selected with a value of 0.000152. So the classification result in sentiment test 1 is Positive. Meanwhile, the sentiment process for test 2 and subsequent ones is the same as for sentiment test 1.

The prediction results from sentiment analysis of tourist reviews in the Lake Toba area using the Naive Bayes method were 31 positive reviews, 378 neutral reviews and 7 negative reviews. Where if there are more positive reviews regarding tourist attractions, especially the Lake Toba area, then tourism processors do not need to improve the tourist attractions much because it is considered that many tourists are satisfied with visiting the Lake Toba tourist attraction and reviews from tourists who are already satisfied can be obtained. influences potential new tourists to visit Lake Toba and does not rule out the possibility of tourists who give positive reviews coming to visit again. However, if there are more negative reviews, then the opposite is true, namely that tourist attractions, especially the Lake Toba area, have a lot of work to do to correct complaints from negative reviews from tourists who have visited..

TESTING

In this process, a system testing stage will be carried out on data that has been collected from social media TikTok, Google From, and TripAdvisor, where the data collected is in the form of comments or reviews from tourists regarding tourist attractions in the Lake Toba area. The data analyzed was 1260 data and divided into training and test data with a ratio of 80:20, so the total training data was 1008 and the test data was 252 data.

Review data from tourist attractions on Lake Toba is labeled first, then text preprocessing is carried out after which it is converted into a numerical representation using TF-IDF (Term Frequency-Inverse Document Frequency). It measures how important a word is in a document based on its frequency in that document compared to the entire corpus. TF-IDF helps identify key words that represent positive, neutral, or negative sentiment in reviews.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import classification_report, accuracy_score
from sklearn.utils import resample

# Membaca data dari file tf_idf_uji.xlsx
file_path = 'tf_df_uji.xlsx'  # Sesuaikan dengan path file Anda
df = pd.read_excel(file_path)

# Asumsikan kolom Term berisi term dan kolom Sentiment berisi label sentimen
terms = df['Term']
sentiments = df['DF']

# Menghapus kelas yang hanya memiliki satu anggota
class_counts = sentiments.value_counts()
to_remove = class_counts[class_counts < 2].index
df_filtered = df[~df['DF'].isin(to_remove)]

# Memperbarui terms dan sentiments
terms = df_filtered['Term']
sentiments = df_filtered['DF']
```

Fig. 2 Naive Bayes sentiment analysis process and accuracy

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

Then, the Naive Bayes MultinomialNB model is used to train data that has been converted into TF-IDF format. Naive Bayes works with the assumption that features (words in this case) are independent of each other, however, in practice it often gives good results in text sentiment analysis.

The model evaluation process is carried out by measuring the accuracy of predictions on test data that has never been seen before. Accuracy gives an idea of how well the model can predict sentiment based on the reviews provided. In addition, the classification report provides deeper insight by showing precision, recall and F1-score for each sentiment class (positive, neutral, negative).

Accuracy: 77.49%
Classification Report:

	precision	recall	f1-score	support
1	1.00	0.00	0.00	43
2	1.00	1.00	1.00	41
3	1.00	1.00	1.00	39
4	0.43	1.00	0.60	32
5	1.00	1.00	1.00	36
accuracy			0.77	191
macro avg	0.89	0.80	0.72	191
weighted avg	0.90	0.77	0.71	191

Fig. 3 Accuracy, Precision, Recall, dan F1-Score

Hasil dari analisis menggunakan metode Naive Bayes pada ulasan tempat wisata di kawasan Danau Toba menunjukkan akurasi sebesar 77.49%. Akurasi ini menggambarkan seberapa baik model dapat memprediksi sentimen berdasarkan ulasan yang diberikan oleh pengunjung.

DISCUSSIONS

In this research, the proposed use of the sentiment analysis method using Naïve Bayes aims to analyze reviews of tourist attractions in the Lake Toba area by giving three class labels, namely positive, neutral and negative. This method compares test text data with training text data to determine the class of tourist reviews of tourist attractions in the Lake Toba area on various platforms such as the TripAdvisor website, TikTok social media or filling out Google Forms. Where the experimental results have predicted results of 31 positive reviews, 378 neutral reviews and 7 negative reviews. So the accuracy results obtained were quite good, namely 77.49% with a total of 1260 data, where the training data was 1008 and the test data was 252 data. The process of analyzing the sentiment of tourist reviews in the Lake Toba area is carried out by cleaning the data or preprocessing the data first. Data preprocessing processes such as case folding, cleaning, tokenization, stopwords and stemming. After preprocessing the data, proceed to the review prediction process using the Naive Bayes method and continue with the process of testing the accuracy of the prediction results using the Naive Bayes method using the confusion matrix.

5. CONCLUSION

Based on the results of testing the sentiment analysis of tourist reviews of the Lake Toba tourist area using Naïve Bayes, it can be concluded that the process of this analysis was carried out by taking review data from several platforms such as the TripAdvisor web, TikTok social media and filling in Google Form with a total of 1260 review data collected. Then the data is labeled sentiment data consisting of positive, neutral and negative review labels. After that, data preprocessing is carried out such as casefolding, cleaning, tokenization, stopwords and stemming. Then the discussion data section becomes training data and test data with a ratio of 80:20 or 1008 training data and 252 test data. Next, a label class prediction process was carried out from tourist reviews of the Lake Toba tourist attraction using the Naïve Bayes method and the prediction results obtained after analysis were 31 positive reviews, 379 neutral reviews and 7 negative reviews with an accuracy result of 77.49%. The prediction results and accuracy of the sentiment analysis that have been carried out are quite effective and good in predicting tourist reviews, so that the results of the semester analysis can be used as a reference for future prospective tourists and can help stakeholders in developing tourism so that it can be better in the future, especially in the region Lake Toba tour.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4287

Submitted : 14 July 2024
Accepted : 25 July 2024
Published : 31 July 2024

6. REFERENCES

Amrullah, A. Z., Sofyan Anas, A., & Hidayat, M. A. J. (2020). Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square. *Jurnal*, *2*(1), 40–44. https://doi.org/10.30812/bite.v2i1.804

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. *Jurnal Teknologi Informasi*, *3*(2), 179. https://doi.org/10.36294/jurti.v3i2.1077

Evita Fitri, Yuri Yuliani, Susy Rosyida, W. G. (2020). Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. *Jurnal Transformatika*, *18*(1), 71–80. https://doi.org/10.26623/transformatika.v18i1.2317

Fitrianah, D., Dwiasnati, S., H, H. H., & Baihaqi, K. A. (2021). Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes. *Faktor Exacta*, *14*(2), 92. https://doi.org/10.30998/faktorexacta.v14i2.9297

Herdhianto, A. (2020). *Sentiment Analysis Menggunakan Naive Bayes Classifier (NBC) pada Tweet tetang Zakat*.

Karyati, I. M. dan C. M. (2020). Analisis Sentimen Terhadap Universitas Gunadarma Berdasarkan Opini Pengguna Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Ilmiah KOMPUTASI*, *19*(4), 507–521.

Luh, N., Sri, W., Ginantra, R., & Wardani, N. W. (2019). IMPLEMENTASI METODA NAÏVE BAYES DAN VECTOR SPACE MODEL DALAM DETEKSI KESAMAAN ARTIKEL JURNAL BERBAHASA. *Jurnal Infomedia*, *4*(2).

Maulani, T. Z., Zulfan, K. S., & Amirullah. (2019). Implementasi Algoritma Naïve Bayes Classifier Dalam Menentukan Topik Tugas Akhir Mahasiswa Berbasis Web. *Jurnal Infomedia*, *4*(1), 33–41.

Nasution, Y. R., & KHAIRUNA, K. (2017). Sistem Pakar Deteksi Awal Penyakit Tuberkulosis Dengan Metode Bayes. *KLOROFIL: Jurnal Ilmu Biologi Dan Terapan*, *1*(1), 17. https://doi.org/10.30821/kfl:jibt.v1i1.1236

Nitamia, M. T., & Februariyanti, H. (2022). Analisis Sentimen Ulasan Ekpedisi J&T Expres Menggunakan Algoritma Naive Bayes. *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, *5*(1), 20–29.

Rahel Lina Simanjuntak, Theresia Romauli Siagian, Vina Anggriani, & Arnita Arnita. (2023). Analisis Sentimen Ulasan Pada Aplikasi E-Commerce Shopee Dengan Menggunakan Algoritma Naïve Bayes. *Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *3*(3), 23–39. https://doi.org/10.55606/teknik.v3i3.2411

Saragih, M. G., Surya, E. D., & B, M. (2021). *Pariwisata Super Prioritas Danau Toba* (S. Widodo (ed.); I, Issue March). Penerbit Andalan. https://www.researchgate.net/profile/Mesra-Mesra/publication/359228856_PARIWISATA_SUPER_PRIORITAS_DANAU_TOBA/links/6230138be32d2203ab413382/PARIWISATA-SUPER-PRIORITAS-DANAU-TOBA.pdf

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Jurnal SIMETRIS*, *10*(2), 681–686.

Sidabutar, A. M., Sarkis, I. M., & Rajagukguk, E. (2023). Analisis Sentimen Cuitan di Terhadap Kawasan Wisata Danau Toba Dengan Metode Naïve Bayes. *Jurnal Ilmiah Teknik Informatika*, *3*(1), 43–53. http://ojs.fikom-methodist.net/index.php/methotika

Toy, K. V. S., Sari, Y. A., & Cholissodin, I. (2021). Analisis Sentimen Twitter menggunakan Metode Naive Bayes dengan Relevance Frequency Feature Selection (Studi Kasus: Opini Masyarakat mengenai Kebijakan New Normal). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(11), 5068–5074. http://j-ptiik.ub.ac.id

* Corresponding author