Sentiment Analysis of Public Opinion Towards Tourism in Bangkalan Regency Using Naïve Bayes Method

Doni Abdul Fatah[1*], *Eka Mala Sari* Rochman[1], *Wahyudi* Setiawan[1], *Ayussy Rahma* Aulia[1], *Fajrul Ihsan* Kamil[1], and *Ahmad* Su'ud[1]

[1]Departmen of Informatics Engineering, Faculty of Engineering, Universitas Trunojoyo Madura, 69162 Bangkalan, Indonesia

Abstract. Sentiment analysis is natural language processing (NLP) that uses text analysis to recognize and extract opinions in text. Analysis is used to convert unstructured information into more structured information, also to determine whether an object has a positive, negative, or neutral tendency, and is an effort to facilitate decision making for tourism managers as a recommendation in developing tourist attractions. In this study, opinions were conducted on tourism reviews in Bangkalan using the Naïve Bayes method. This method is a machine learning algorithm to classify text into concepts that are easy to understand and provide accurate results with high efficiency. This method is proven to provide excellent results with a high level of accuracy, especially for large data, but has some drawbacks, sensitive to feature selection. Thus, a feature selection process is needed to improve classification efficiency by reducing the amount of data analyzed, with the Information Gain feature selection method. The word weighting method uses TF-IDF, while the data used comes from google maps reviews taken through web scraping, where tourist visitors provide reviews and ratings of places that have been visited. However, the large number of reviews can make it difficult for tourist attractions managers to manage them, so the process of labeling the sentiment class of the review data obtained 3649 reviews, with 2583 positive, 275 negative, and 457 neutral. Based on the test results that have been carried out using the Information Gain threshold of 0.0001, 0.0003, and 0.0007 can improve the accuracy of the Naïve Bayes model, for the best test at threshold 0.0007, with an accuracy value of 78.68%, precision 80.44%, recall 82.59%, and f1-score 82.53%, from the test results it shows that the use of information gain feature selection and SMOTE technique has a fairly good performance in classifying public opinion sentiment data on tourism in Bangkalan Regency, meaning that tourism management is good seen from the results of visitor satisfaction sentiment.

Keywords: Information Gain, Naïve Bayes, Sentiment Analysis, Tourism.

1 Introduction

Tourism is one of the attractions that must be developed in each region because the growth of this sector can increase opportunities to provide employment and business for the surrounding people, which in turn can encourage regional development and regional income [1]. Bangkalan Regency is one of the areas in Madura that has a lot of tourism potential that can be developed [2][3], whether from natural, religious, or artificial tourism. The Bangkalan Regency Government has made a tourism policy that emphasizes the arrangement of tourist attractions to attract tourists and improve existing tourist attractions in the area. With this policy, existing tourist attractions in the area need to be developed [3], to find out whether the success of tourism in an area can be measured by increasing the number of tourists who come to the tourist attractions [4]. However, tourism managers must know what affects the satisfaction of visitors or tourists. So a study is needed to find out, one of which is by knowing the opinions of visitors regarding these tourist attractions.

Sentiment analysis is a way to find out the opinion of user satisfaction with a tour or system, besides that it can also be used to measure whether visitors to tourist attractions like the tourist attractions or not[5]. Sentiment analysis is a part of natural language processing (NLP) that uses text analysis to recognize and extract opinions from text, it is used to convert unstructured information into more structured information and is also used to determine whether an object tends to be positive, negative, or neutral [6]. There are many ways to perform topic sentiment analysis, one of which is using machine learning [7].

Machine learning is a technology that allows machines to learn data patterns and perform certain tasks independently [7]. Machine learning in sentiment analysis is used to find sentiment patterns in text. This allows machine learning to be used to classify text based on positive, negative, or neutral sentiment. The Naive

[*] Corresponding author: doni.fatah@trunojoyo.ac.id

Bayes method is a fairly popular machine learning method or algorithm used to perform text classification, providing accurate results and high efficiency, but the naïve Bayes method still has some drawbacks when used, one of which is sensitive to feature selection. Classification performance can be reduced by a very large number of features. As a result, a feature selection process is required to make the classification process more efficient by reducing the amount of data to be analyzed. So in this study, the novelty is to use a feature selection method, namely Information gain, which is to overcome the shortcomings that exist in the naïve Bayes method in feature selection, Information gain is a method that works by selecting features with the highest weight according to the desired number of features. This method uses entropy to find the best term. The greater the Information Gain value of a term, the more significant the feature is. Meanwhile, to measure how often a word appears in a document, and count the number of documents in which the word appears, the Term Frequency-Inverse Document Frequency (TF-IDF) word weighting is used [8].

The Naïve Bayes method in this study uses increased TF-IDF word weighting and Information Gain feature selection, where this method is proven to have very good results with a very high level of data accuracy [9][10]. The data source in this study comes from review data obtained through reviews of tourist visitors in Bangkalan Regency which are taken from Google Maps reviews, because from Google Maps reviews it can be seen that visitors' opinions on the tour and also know the rating of places that have been visited [11]. In Bangkalan Regency, many visitors to tourist attractions have provided reviews on the Google Maps application, which can help tourism managers know the level of visitor satisfaction through the reviews given [12]. However, the large number of reviews can make it difficult for tourism managers to manage them efficiently. Therefore, to process data derived from tourist visitor reviews in Bangkalan Regency taken from Google Maps reviews, naïve Bayes is used to calculate probabilities. One of the advantages of the naïve Bayes algorithm when compared to other machine learning algorithms is the high level of accuracy, making the naïve Bayes algorithm suitable for use on very large data [13].

The review data in this study, which had previously been obtained through reviews from Google Maps, had an unbalanced number of classes in the tourism review data. So it is necessary to do a data balancing process, another addition to this research is the use of sampling techniques to overcome data balancing. The process of handling a balanced dataset in classification can be done instantly without following certain steps. However, when the number of dataset classes is unbalanced, sampling techniques are needed to handle unbalanced datasets so that the classification model continues to function properly. Thus, the novelty in this research, using SMOTE (Synthetic Minority Oversampling Technique) is used to overcome the imbalance of classes in the dataset in this research problem [14].

The ideas in this research that have been put forward above are compared with several similar studies that have been conducted by several other researchers including, research to analyze the sentiment of the Bjorka hacker on Twitter social media by comparing the accuracy results of the Multinomial Naïve Bayes algorithm, Naïve Bayes Bernoulli, and Naïve Bayes Gaussian, using 1000 data, with TF-IDF word weighting showing the Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes algorithms have 73% accuracy, 73% precision, and 100% recall; Bernoulli Naïve Bayes has 72% accuracy, 73% precision, and 98% recall; and Gaussian Naïve Bayes has 55% accuracy, 75% precision, and 63% recall [15]. In similar research, the use of the Support Vector Machine (SVM) method is used to conduct sentiment analysis on Madura tourism, where the data used comes from various social media platforms, that discuss Madura tourism. Reviews of public opinion are divided into three categories positive, negative, and neutral, from the test results using K = 5 fold cross validation resulted in a positive sentiment of 192 tweets and an accuracy of 92.592% using Confusion Matrix [16]. Another advanced research, comparing the Naïve Bayes algorithm with SVM on Twitter data. The analysis results show that the Naïve bayes algorithm can achieve 3.45% accuracy, precision 0.02, recall 0.04, and f1-score 0.03 compared to the sentiment analysis model using a Support Vector Machine. The data used in this study amounted to 2030 data and were divided into two models, namely training model data as much as 1624 data, and test data totaling 406 data [17]. Research on the use of SMOTE techniques to overcome unbalanced data. The secondary data used in this study came from the national socioeconomic survey, which consisted of 494 samples consisting of 8 independent variables and 1 dependent variable. The CART (Classification and Regression Trees) method was used [18]. This study shows that the model with SMOTE produces more accurate values compared to the model without SMOTE. The model with SMOTE produced a higher sensitivity of 67.05% compared to the previous value of only 36.36% [19] [20].

Based on a series of explanations that have been explained previously in this study, namely Sentiment Analysis of Public Opinion Towards Tourism in Bangkalan Regency Using the Naïve Bayes Method to overcome the weaknesses of the Naïve Bayes method with the addition of Information gain for feature selection, TF-IDF for word weighting and using SMOTE to overcome class imbalance in the dataset. It is hoped that these additions can provide a high level of accuracy in providing the results of tourism visitor reviews in Bangkalan district.

2 Methods

The research method used to describe the research design carried out to complete research on sentiment analysis of public opinion towards tourism in Bangkalan

Regency using the Naïve Bayes method, which is explained according to the figure 1 below:

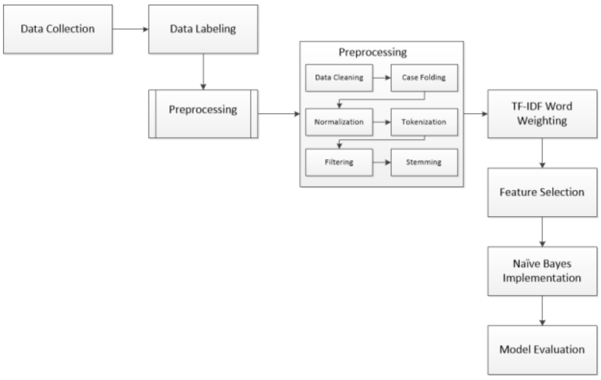

Fig. 1. Stages of the research.

In the picture above is an overview of the process carried out in this study.

2.1 Data collection and labeling

The data used in this research is secondary data derived from tourist visitor reviews on Google Maps. The selected tourist destinations are 10 recommended tours in Bangkalan Regency, namely Bukit Pelalangan Arosbaya, Bukit Geger, Bukit Jaddih, Siring Kemuning Beach, Labuhan Mangrove Education Park, Rongkang Beach, Sambilangan Lighthouse, Syaikhona Kholil Tomb, Beramah Lantern Hill, and Sumber Pocong. Data collection is done by web scraping using webharvy tools with a total of 3649 review data from 2021 to 2023.

Sentiment reviews are labeled as positive, neutral, or negative, based on visitor stars. 1 and 2 stars are categorized as negative sentiment, 3 stars as neutral sentiment, and 4 and 5 stars as positive[21][22].

2.2 Data preprocessing

The preprocessing process to produce clean data the stages used are:

a) Data Cleaning is a process intended to ensure that the data from the data set has the best accuracy, consistency, and usability.

b) Case Folding is used to homogenize or convert all letters from "a" to "z" in the dataset into lowercase letters.

c) Normalization is a procedure to equalize or homogenize words that are written in different ways but have the same meaning as well as converting nonstandard words into standard words. This is done by using a slangword dictionary obtained from the Indonesian Colloquial Lexicon which can be accessed on Github, as well as a slangword dictionary created by the author based on the needs of the dataset.

d) Tokenization is by breaking the sentence into words or tokens. The tokenization stage is performed using the split() method in the python programming language. This method breaks the words from each sentence into tokens.

e) Filtering is the process of removing a list of unimportant and useless words using the stopword word list from the nlp_id library.

f) Stemming, at this stage, words with affixes are converted into basic words. For the stemming stage, a standard word dictionary is needed, in this study using the Sastrawi library using the StemmerFactory module, from the preprocessing process getting the final data amount of 3326 clean review data.

After the data is cleaned, TF-IDF word weighting will be carried out, where this method combines two concepts, namely Term Frequency and Document Frequency.

2.3 TF-IDF word weighting

After preprocessing, the dataset must be extracted first. Because the data contained in the dataset of tourist visitor reviews is data in the form of text or categorical data and machine learning cannot accept input in the form of categorical data, therefore feature extraction is needed to convert categorical data into numerical data, namely by using the TF IDF (Term Frequency Inverse Document Frequency) word weighting algorithm [23][24]. TF-IDF will give weight to each word, where this method combines two concepts, namely Term Frequency and Document Frequency. Term Frequency measures how often a word appears in a document, while for Document Frequency to find out the number of documents in which a word appears, and the less the frequency of occurrence [25], the lower the weight value, according to the following formula:

$$\mathrm{t}f_{t,d} = \frac{N_{t,d}}{N_d} \tag{1}$$

$$idf_t = log^{n}_{nk} \tag{2}$$

$$tf - idf_t = tf_{t,d} * \ idf_t \tag{3}$$

Where, tft,d is the term frequency value t of document d. Nt,d is the appearance of term t in document d. Nd is the total terms in document d. $idft$ is the idf value of term t. n is the number of document collections. nk is the number of documents that contain term t.

2.4 Oversampling data

The next stage is Oversampling this data has the aim of dealing with class imbalances in the dataset used in the study, because there are differences in the number of classes in the dataset related to the data used in this study [26], so that the data is balanced using SMOTE (Synthetic Minority Over-sampling Technique) oversampling. The SMOTE technique is performed by duplicating samples from minority classes so as to generate new synthetic samples through extrapolation of existing minority samples using random samples. By applying SMOTE to unbalanced data, the performance of the model in predicting minority classes can be improved [27], using the following equation:

$$Dist = \sqrt{(X_1 - X_1)^2 + (X_2 - X_2)^2 + \cdots + (X_n - X_n)^2} \qquad (4)$$

$$Xsyn = Xi + (Xknn - Xi)\, x\, \delta \qquad (5)$$

Where, *Dist* is the euclidean distance. *Xn* is the nth attribute value. *Xsyn* is synthetic data created to generate new data. *Xi* is the data to be replicated. *Xknn* is the data that has the closest distance from the data to be replicated. δ is a random number between 0 and 1.

2.5 Feature selection

Next is feature selection using information gain, which is very important in performing text classification by forming a vector space to improve scalability, efficiency, and accuracy in the text clustering process. Information gain works by selecting features that have the highest weight according to the desired number of features. Information gain involves entropy to find the best term. If the information gain value of a term is greater, then the feature is considered more significant and more important [28].

2.6 Naïve bayes implementation

The next stage of sentiment analysis, training data is used for model training using the Naïve Bayes method so that a model for classification is obtained. Naïve Bayes Classifier is a probabilistic algorithm used to predict a situation [18]. In its use, it uses the concept of probability theory which involves predicting the likelihood of a future event based on data that has been collected in the past. As a probability concept, Naïve Bayes Classifier can be used to classify text documents into certain classes with high accuracy and is able to process large amounts of data [12]. In classifying text documents, there are several processes that must be done, namely, finding the probability of each document category, finding the probability of occurrence of each word in each document category, determining the category of documents to be classified based on calculations from the first and second stages.

The next stage of sentiment analysis, training data is used for model training using the Naïve Bayes method so that a model for classification is obtained [29]. The Naïve Bayes Classifier algorithm assigns a target value to new data using the $Vmap$ value, which is the highest possible value of all members of the domain set V [30]. Each review data is represented with attribute pairs $x1$, $x2$, $x3$ ... . . xn where $x1$ is the first word, $x2$ is the second word and so on. Whereas V is the set of sentiment categories [31]. During classification, the algorithm will look for the highest probability of all tested categories ($Vmap$), where the equation is as follows:

$$Vmap = \frac{arg\ max_{P\ (x_1,x_2,x_3\ \dots\ x_n|V_j)P(V_j)}}{V_{j\ \ eV}{}^{P\ (x_1,x_2,x_3\ \dots\ x_n)}} \qquad (6)$$

Where, Vj is the review category j = 1,2,3, ...n Where in this study: $j1$ is a positive review category, $j2$ is a negative review category, $j3$ is a neutral review category. $P\ (xi|Vj\ )$ is the probability of xi in category Vj. $P(Vj)$ is the probability of Vj.

2.7 Model evaluation

After the model is completed, the last stage is testing and evaluation. This stage will measure accuracy using a confusion matrix to compare actual categories and predicted categories. This measurement is done by calculating the accuracy, precision, recall, and f1-score values. In multi-class classification evaluation, there are 3 commonly used matrices, namely, accuracy, precision, and recall. Accuracy measures the ratio of correct predictions to the total data evaluated. Precision measures the level of accuracy between the requested data and the answer or result provided by the system. Meanwhile, recall is used to measure the amount of data that is correctly classified against the total data that should belong to that class [32].

At this stage, data testing on the model will be carried out by evaluating the extent of naïve Bayes performance on the tourist visitor review dataset using information gain feature selection with threshold values of 0.0001, 0.0003, and 0.0007 to explore the effect of using thresholds in selecting features that can affect model performance and features that have weight values below the three threshold values will not be used. Information gain has an important role as a feature selector so that the accuracy of the system can be better. The selected feature is a feature with an information gain weight that has a value not equal to zero and the feature will be selected using a threshold to produce an output in the form of the best feature [33].

The stages carried out in this research start from data collection, data labeling, preprocessing, TF-IDF word weighting, Feature selection, naive Bayes implementation, and evaluation model. From this process, the results of this study are obtained in the form of the best accuracy, precision, recall, and F-1 score values based on the results of the test scenario using the confusion matrix.

3 Result and discussion

The data used amounted to 3649 items, consisting of 2583 positive reviews, 275 negative reviews, and 457 neutral reviews. Furthermore, the preprocessing stage was carried out and obtained clean data as well as 3326 review data. After that, TF-IDF word weighting is carried out. The data is then balanced by the number of sentiments in the reviews using the SMOTE technique. After the data balancing process is carried out with the SMOTE technique, the number of positive, negative, and neutral reviews is 2583. The data is divided into training data and testing data, and feature selection is carried out using information gain. In the sentiment analysis stage, training data is used to train the model with the multinomial Naïve Bayes method for classification. Once the model is ready, testing and evaluation are done using the testing data to measure accuracy.

3.1 Data collection

To collect data for this research, w[11]e used data scraping techniques from reviews of tourist attractions in Bangkalan Regency. WebHarvy software is used to collect data on WebHarvy by retrieving data from Google Maps, such as name, time, stars, and reviews on the website.

3.2 Data labeling

Tourist review data is labeled based on the stars obtained by determining each class of visitor review data. Each class of each tourist visitor review data, whether it belongs to positive, negative, or neutral sentiment. Number of labeling results shown in Table 1.

Table 1. Number of labeling results.

Positive	Negative	Neutral
2583	275	457

Based on the table above, it can be seen that the number of positive labels is 2583 data, negative labels are 275 data, and neutral labels are 457 data.

3.3 Preprocessing

a. Data cleaning

Reviews	Data Cleaning
Exotic former limestone mining is pretty good for eye wash and a good spot for photos; it's just that the road access to the location is still not good.	Exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good

b. Casefolding

Data Cleaning	Casefolding
Exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good	exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good.

c. Normalization

Casefolding	Normalization
exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good.	exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good.

d. Tokenization

Normalisasi	Tokenization
exotic former limestone mining is pretty good for eye wash good spot for it's just that the access road to the location is still not good.	['former', 'mining', 'stone', 'limestone', 'which', 'exotic', 'not bad', 'good', 'make', 'wash', 'eye', 'spot', 'which', 'good', 'make', 'only', 'just', 'access', 'road', 'to', 'location', 'still', 'not yet', 'good']

e. Filtering

Tokenization	Filtering
['former', 'mining', 'stone', 'limestone', 'which', 'exotic', 'not bad', 'good', 'make', 'wash', 'eye', 'spot', 'which', 'good', 'make', 'only', 'just', 'access', 'road', 'to', 'location', 'still', 'not yet', 'good']	['former', 'mining', 'stone', 'limestone', 'exotic', 'not bad', 'good', 'wash', 'eye', 'spot', 'nice', 'access', 'road', 'location', 'not yet', 'good']

f. Stemming

Filtering	Stemming
['former', 'mining', 'stone', 'limestone', 'exotic', 'not bad', 'good', 'wash', 'eye', 'spot', 'nice', 'access', 'road', 'location', 'not yet', 'good']	['used', 'quarry', 'stone', 'limestone', 'exotic', 'not bad', 'good', 'wash', 'eye', 'spot', 'nice', 'access', 'road', 'location', 'not yet', 'good']

Each stage of text preprocessing affects the number of words contained in the review text. Table 2 is the change in the number of words after the text preprocessing stages. Changes in text preprocessing word count shown in Table 2.

Table 2. Changes in text preprocessing word count.

Category	Word Count
Reviews	59516
Data Cleaning	57812
Case Folding	57812
Normalization	57873
Tokenization	57883
Filtering	36348
Stemming	36348

3.4 TF-IDF word weighting

The TF-IDF algorithm calculates the importance of each word in a document by assigning a weight to each word after performing data division. This method can be used to convert each word in a text document into a numerical frequency. By using the scikit-learn library, TF-IDF word weighting can be done in Python by using the TfidfVectorizer() function. The results can be seen in the table below.

Table 3. Results of TF-IDF word weighting.

Term	TF-IDF Weight
Good	217.261702
Photo	98.819628
Travel	98.223616
Log in	91.978054

Term	TF-IDF Weight
Road	89.9065565
Mikat	0.038767
Warawiri	0.038767
Upload	0.038767
Published	0.038767
Dawn	0.038767

3.5 Naïve bayes implementation

The naive bayes classification process is carried out on data that has gone through the tf-idf weighting stage and information gain feature selection using training data with predictions using testing data. From the classification results, the accuracy, precision, recall and F1-score results are obtained. The library used for the naive bayes classification process is scikit-learn. The Scikit-learn library used is MultinomialNB. The following is the implementation of naïve bayes is shown in Sourcode 1.

```
from     sklearn.naive_bayes     import
MultinomialNB

# Training model Naive Bayes Threshold
0.0003
naive_bayes1   =   MultinomialNB(alpha   =
0.1,                      fit_prior=True)
naive_bayes1.fit(X_train_info_gain1,
y_train)

#Testing Model
y_pred_gain1                            =
naive_bayes1.predict(X_test_info_gain1)
```

Sourcode 1. Implementation of naïve bayes.

3.6 Model evaluation

In this scenario, the Naïve Bayes algorithm is tested using the Information Gain feature selection and SMOTE technique. In this test, the Information Gain feature selection and SMOTE technique are carried out before classifying using the Naïve Bayes algorithm. Evaluation is done using confusion matrix to see the accuracy results of Naïve Bayes algorithm using information gain with threshold 0.0001, 0.0003, and 0.0005. The results of the testing scenario are shown in the table below. Description in table 4, NB: Naïve Bayes, IG: Information Gain.

Table 4. Comparison of Naïve Bayes test results with SMOTE.

Model	Confussion Matrix In (%)					
	Number of Features	Time	Accuracy	Precision	Recall	F1 score
NB + IG threshold 0.0001 + SMOTE	4094	0.1242	82.62%	83.06%	82.20%	82.14%
NB + IG threshold 0.0003 + SMOTE	2349	0.0402	82.68%	82.51%	82.49%	82.44%
NB + IG threshold 0.0007 + SMOTE	1906	0.0226	80.89%	80.65%	80.78%	80.70%

Based on the table, it can be seen that the accuracy value obtained from each model by applying SMOTE oversampling. The evaluation results of the Naïve Bayes method using Information Gain feature selection at a threshold of 0.0001 with the number of features is 4094 with time 0.1242 resulted in an accuracy of 82.62%, precision 83.06%, recall 82.20% and f1-score 82.14%, while information gain with a threshold of 0. 0003 with the number of features is 2349 with time 0.0402 resulting in accuracy of 82.68%, precision 82.51%, recall 82.49%, and f1-score 82.44% and threshold 0.0007 with the number of features is 1906 with time 0.0226 resulting in an accuracy value of 80.89%, precision 80.65%, recall 80.78% and f1-score 80.70%.
The results of processing tourist visitor reviews in Bangkalan Regency using SMOTE produce the best accuracy, namely naïve bayes with an information gain threshold of 0.0002, with a total positive sentiment of 2597, with positive words that most often appear are the words "good", "tour", "hill", "photo", and "place". This shows that visitors have a good and pleasant experience in visiting tourism in Bangkalan Regency. With a total negative sentiment of 274, with negative sentiment the words that most often appear are the words "enter", "extortion", "parakeet", "not", and "road". As for the number of neutral sentiments of 455, with neutral sentiments the words that appear most often are the words "place", "good", "beautiful", "photo", and "cool".

3.7 Analysis result

The results of the comparison of the accuracy, precision, recall, and f1-score results of the naïve Bayes model and information gained with SMOTE in the figure below.

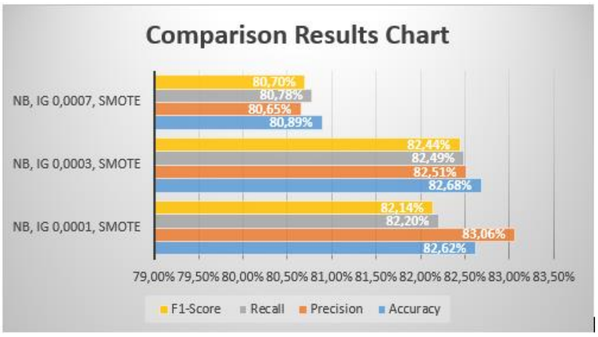

Fig. 2. Naïve bayes comparison results chart.

Based on the test results in this study, it shows that the use of SMOTE increases the accuracy of the Naïve

Bayes model to 82.81% with balanced precision and recall. The application of Information Gain with thresholds 0.0001, 0.0003, and 0.0007 also increases the accuracy of the Naïve Bayes model to 78.37%, 78.67%, and 78.67%, respectively. Using a combination of the Naïve Bayes algorithm with an information gain threshold of 0.0003 and SMOTE produces the best accuracy of 82.68%. This method helps classify sentiment more precisely and efficiently, providing deep insight into visitor satisfaction with tourism in Bangkalan Regency. The use of information gain with the right threshold and SMOTE is the best choice for analyzing visitor reviews in Bangkalan Regency.

4 Conclusion

Evaluation of the results of using SMOTE in the Naïve Bayes method using Information Gain feature selection at a threshold of 0.0001 resulted in an accuracy of 82.62%, precision 83.06%, recall 82.20%, and f1-score 82.14%, while information gain with a threshold of 0.0003 resulted in an accuracy of 82.68%, precision 82.51%, recall 82.49%, and f1-score 82.44%, and threshold 0.0007 resulted in an accuracy value of 80.89%, precision 80.65%, recall 80.78%, and f1-score 80.70%. From the test results, it shows that the threshold value of 0.0003 gets the best accuracy value from the other thresholds. In other words, Naïve Bayes optimization using information gain and SMOTE feature selection gives better results in analyzing tourist visitor reviews.

Based on this research, it can be concluded that the use of information gain with the selection of the right threshold value and the use of SMOTE produce good accuracy. With high accuracy and good precision, recall, and f1-score values, this method helps in classifying sentiments more precisely and efficiently. Thus, the Naïve Bayes model optimized with SMOTE and information gain techniques is a better choice for the analysis of tourist visitor reviews in Bangkalan Regency and can provide deeper insight into visitor satisfaction with tourism in Bangkalan Regency.

References

1. N. Nurhayati, “ The Impact of Tourism Village Development on Community Welfare,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.
2. L. Sri and L. S. W. Wulandari, “ The potential of bangkalan regency as the center of Madura island tourism which has a strategic location and beautiful karst landscape,” 2020.
3. One-Stop Investment and Integrated Services Office, "Potential of the industrial sector,” 2023. http://investment.bangkalankab.go.id/pontensi_unggulan (accessed Nov. 10, 2023).
4. J. Zhou, Y. Lu, H.-N. Dai, H. Wang, and H. Xiao, “Sentiment Analysis of Chinese Microblog Based on Stacked Bidirectional LSTM,” *IEEE Access*, vol. 7, pp. 38856–38866, 2019.
5. A. Imron, “ Sentiment Analysis of Tourist Attractions in Rembang Regency Using Naive Bayes Classifier Method,” *Tek. Inform.*, pp. 10–13, 2019, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/14268.
6. D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.
7. S. B. Kotsiantis, “Erratum: Feature selection for machine learning classification problems: A recent overview (Artificial Intelligence Review (2011)),” *Artif. Intell. Rev.*, vol. 42, no. 1, p. 157, 2014.
8. S. Aggarwal and N. Chugh, “Correction to: Review of Machine Learning Techniques for EEG Based Brain Computer Interface (Archives of Computational Methods in Engineering, (2022), 29, 5, (3001-3020), 10.1007/s11831-021-09684-6),” *Arch. Comput. Methods Eng.*, vol. 29, no. 5, p. 3531, 2022.
9. S. Ruan, B. Chen, K. Song, and H. Li, “Weighted naïve Bayes text classification algorithm based on improved distance correlation coefficient,” *Neural Comput. Appl.*, vol. 34, no. 4, pp. 2729–2738, 2022.
10. E. Redivo, C. Viroli, and A. Farcomeni, “Quantile-distribution functions and their use for classification, with application to naïve Bayes classifiers,” *Stat. Comput.*, vol. 33, no. 2, p. 55, 2023.
11. J. Singh, G. Singh, and R. Singh, “Optimization of sentiment analysis using machine learning classifiers,” *Human-centric Comput. Inf. Sci.*, vol. 7, no. 1, p. 32, 2017.
12. F. Itoo, Meenakshi, and S. Singh, “Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection,” *Int. J. Inf. Technol.*, vol. 13, no. 4, pp. 1503–1511, 2021.
13. H. Langseth and T. D. Nielsen, “Classification using Hierarchical Naïve Bayes models,” *Mach. Learn.*, vol. 63, no. 2, pp. 135–159, 2006.
14. R. A. Barro, I. D. Sulvianti, and F. M. Afendi, “ Application of Synthetic Minority Oversampling Technique (SMOTE) to Unbalanced Data in Herbal Composition Modeling,” *J. Stat.*, vol. 1, no. 1, pp. 1–6, 2013, [Online]. Available: https://doi.org/10.29244/xplore.v1i1.12424.
15. H. Ashari, D. Arifianto, H. Azizah, and A. Faruq, “ Performance Comparison of Multinomial Naïve Bayes Algorithm (MNB), Multivariate Bernoulli and Rocchio Algorithm in Classification of Indonesian Language Hoax News Content on Social Media,” *Http://Repository.Unmuhjember.Ac.Id*, pp. 1–12, 2020.
16. Doni Abdul Fatah, Eka Mala Sari Rochman, Fajrul Ihsan Kamil, and Ahmad Su’ud, “Sentiment Analysis of Madura Tourism Opinion Using Support Vector Machine (SVM),” *Tech. Rom. J.*

Appl. Sci. Technol., vol. 16, pp. 243–249, Oct. 2023.

17. Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, " Comparative Analysis of Data Mining Classification Algorithms for Blogger Dataset with Rapid Miner," *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 20–28, 2018.
18. I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, "General Performance Score for classification problems," *Appl. Intell.*, vol. 52, no. 10, pp. 12049–12063, 2022.
19. M. I. Fikri, T. S. Sabrila, and Y. Azhar, " Comparison of Naïve Bayes and Support Vector Machine Methods on Twitter Sentiment Analysis," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.
20. M. I. Putri and I. Kharisudin, " Application of Synthetic Minority Oversampling Technique (SMOTE) to Sentiment Analysis of Tokopedia Marketplace Application User Review Data," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 759–766, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/.
21. R. Chan *et al.*, "What should AI see? Using the public's opinion to determine the perception of an AI," *AI Ethics*, vol. 3, no. 4, pp. 1381–1405, 2023.
22. J. Li, H. Sun, and J. Li, "Beyond confusion matrix: learning from multiple annotators with awareness of instance features," *Mach. Learn.*, vol. 112, no. 3, pp. 1053–1075, 2023.
23. S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 30, 2019.
24. T.-P. Hong, C.-W. Lin, K.-T. Yang, and S.-L. Wang, "Using TF-IDF to hide sensitive itemsets," *Appl. Intell.*, vol. 38, no. 4, pp. 502–510, 2013.
25. A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 3007–3054, 2021.
26. Q. Dai, J. Liu, and J.-L. Zhao, "Distance-based arranging oversampling technique for imbalanced data," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1323–1342, 2023.
27. S. Guo, Y. Liu, R. Chen, X. Sun, and X. Wang, "Improved SMOTE Algorithm to Deal with Imbalanced Activity Classes in Smart Homes," *Neural Process. Lett.*, vol. 50, no. 2, pp. 1503–1526, 2019.
28. D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, 2023.
29. X. Chen, Y. Xue, H. Zhao, X. Lu, X. Hu, and Z. Ma, "A novel feature extraction methodology for sentiment analysis of product reviews," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6625–6642, 2019.
30. M. S. Vural and M. Gök, "Criminal prediction using Naive Bayes theory," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2581–2592, 2017.
31. Z.-L. Xiang, X.-R. Yu, and D.-K. Kang, "Experimental analysis of naïve Bayes classifier based on an attribute weighting framework with smooth kernel density estimations," *Appl. Intell.*, vol. 44, no. 3, pp. 611–620, 2016.
32. J. Asbee, K. Kelly, T. McMahan, and T. D. Parsons, "Machine learning classification analysis for an adaptive virtual reality Stroop task," *Virtual Real.*, vol. 27, no. 2, pp. 1391–1407, 2023,.
33. S.-H. Park and J. Fürnkranz, "Efficient implementation of class-based decomposition schemes for Naïve Bayes," *Mach. Learn.*, vol. 96, no. 3, pp. 295–309, 2014.