Proceeding of International Conference on Digital Advance Tourism, Management and Technology 2023
Volume. 1 No. 1 2023 e-ISSN : 3031-5999, Hal 215-230
DOI: https://doi.org/10.56910/ictmt.v1i1.65

Comparative Analysis Of Performance Levels Of Svm And Naïve Bayes Algorithm For Lifestyle Classification On Twitter Social Media

Fadlila Nurwanda[1], Winita Sulandari[2*], Yuliana Susanti[3], Zakya Reyhana[4]
[1-3] Department of Staistics, Universitas Sebelas Maret, Surakarta, Indonesia
[4]Solusi247, Yogyakarta, Indonesia

*Corresponding email: *winita@mipa.uns.ac.id ; fadlilanurwanda@gmail.com*

Abstract. *Lifestyle is how individuals express themselves through their activities and interests and utilize their financial resources and available time. Twitter is a social network platform that allows people to express opinions and directly criticize various topics, including the recently widely discussed lifestyle topics. Topic classification on Twitter is central in facilitating the search, recommendation, and management of relevant content for users. This research aims to analyze public sentiment regarding lifestyle using 11,000 pieces of data with the keywords "concert", "watching films", "smoking", and others related to lifestyle. Research data is labeled according to the sentiment of public opinion towards lifestyle. Negative polarity for data that has the context of "underestimating", "insulting", "sarcastic", and "feeling sad". Positive polarity for data that has the context of "grateful", "praying", "feeling happy", and "encouraging". Neutral polarity for data that has the contexts "ask", "predict", and "feel surprised". Next, the data enters the pre-processing stage, which consists of case-folding, tokenization, stopword removal, and lemmatizing. The analysis continues by dividing the data into training and test data with a ratio of 70%:30%. Sentiment analysis uses an algorithm Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC). The analysis results show that the SVM algorithm provides better classification than NBC. In this case, the SVM algorithm produces accuracy, precision, recall, and value F1-Score the same, namely 61%.*

Keywords: *Sentiment Analysis; Naïve Bayes Classifier; Support Vector Machine; Lifestyle.*

INTRODUCTION

Lifestyle is the way individuals express themselves through their activities and interests, and how they utilize their financial resources and available time. This lifestyle develops through the influence of social interaction (Article, 2006). Some Twitter users are active in the music community and often talk about concerts, music festivals or performances live their favorite artists. Users with an interest in automotive may talk about cars, motorbikes, racing, or developments in the automotive industry. Gamers often share their experiences in playing video games. Users interested in food, diet, and healthy lifestyles may share recipes, discuss diet trends, or discuss organic foods and special diet options. Many Twitter users also talk about work, careers, and professional development. There are communities on Twitter that focus on film and television. The interesting thing about Twitter is its ability to allow users to engage in conversations and communities that align with their interests.

Twitter has become one of the favorite platforms among other social media because it gives users the ability to participate in dynamic global conversations (Anger & Kittl, 2011). Twitter is also an active place for discussion and exchange of information. The exchange of information on Twitter is divided into various topic classifications that reflect a variety of interests, issues, and conversations covering everything from economics and industry to

Received November 19, 2023; Revised Desember 01 , 2023; Accepted Desember 31, 2023
* Winita Sulandari , *winita@mipa.uns.ac.id*

lifestyle, politics, and legal violations. Classification in social media refers to the process of grouping or categorizing content, users, or information based on certain attributes or characteristics. In some cases, conversations about lifestyle on Twitter can influence public perceptions about a particular topic, thereby influencing individual behavior.

Twitter is platform a social network that facilitates people to express opinions and provide direct criticism on various topics, including lifestyle topics that are widely discussed by Indonesian society. This results in pros and cons in the form of suggestions and complaints. The large number of public opinions regarding lifestyle topics consisting of positive and negative comments gives certain concerns. As a result, action is needed in the form of sentiment analysis to evaluate people's opinions regarding lifestyle, by grouping comments into positive, neutral, and negative categories (Amelia et al., 2022). For example, if there is a tweet that contains the word praise, then the tweet will be given positive polarity, and tweets containing elements of hatred will be given a negative polarity.

Topic classification on Twitter has a central role in facilitating the search, recommendation, and management of relevant content for users. This allows Twitter to serve content tailored to individual interests, provide account recommendations and trending relevant topics, as well as helping companies identify potential target markets. Then, classification also plays an important role in filtering inappropriate content, securing the environment online, as well as making it easier for users to engage in trending conversations. Twitter's classification features create a more focused, informative, and safe user experience.

Based on Dixon (2022), the number of Twitter users in Indonesia is ranked fifth in the world. Amelia et al., (2022) explained that the Twitter application can be used to gain insight into how people view and feel about current lifestyles. As many as 18.45 million Indonesians actively use Twitter, this is a potential data source for understanding people's views. These abundances of data would make sentiment analysis a relevant option for extracting insights from data. Sentiment analysis is a research domain that focuses on observing a person's views, feelings, judgments, actions, and emotions based on the use of written language.

The application of sentiment analysis systems has become common in various business sectors and social environments because individual views are at the core of almost all human actions and are the main factor in shaping personal behavior (Liu, 2012). The main focus of sentiment analysis is to categorize the author's reaction to a topic into three main classifications, namely positive, negative, and neutral (Beigi et al., 2016). The term "opinion mining" can refer to sentiment analysis, as both describe similar research domains and can be considered part of the sub-field of subjectivity analysis (Pang & Lee, 2008).

Public sentiment analysis research uses classification methods such as algorithms Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC) (Naufal et al., 2023) and (Fitri, 2020) have been implemented. SVM utilizes a hyperplane to divide the sample into two groups, taking into account the hyperplane with distance margin maximum (maximum-margin hyperplane), which is measured from the hyperplane to the nearest vector. Meanwhile, NBC operates by assuming that the existence of a feature in one class is not dependent on the existence of features in other classes.

The NBC algorithm has several advantages, namely that it has been proven to have high speed and accuracy when used in large data databases. In addition, the NBC classification method is one of the most popular text classification methods and has good performance. NBC algorithm sentiment analysis has been used in several related studies (Amelia et al., 2022). One of the advantages of the SVM algorithm is that it has kernels that can implement high-dimensional non-linear input separation, some of the kernels contained in SVM are Radial Basis Function (RBF), polynomial, and linear (Ahmad & Gata, 2022).

Considering the advantages of the two methods, i.e, NBC and SVM algorithms, this study focuses on their implementation to examine lifestyle classification on Twitter social media and compare the result. The results of this research are expected to provide benefits for the realm of statistics and benefits for related media. For the field of statistics, the benefits lie primarily in the progress and implementation of statistics in the field of data mining, especially in text mining for sentiment analysis regarding current lifestyle trends. Apart from that, the benefits are also perceived by the media involved, because it helps them understand the public's reaction to the programs they present. This will provide valuable information for future consideration and decision-making.

METHOD

The following are the stages carried out in this research.

1. **Data collection**

   Data collection on Twitter social media is a way of crawling according to the specified topic, namely lifestyle topics with the keywords "concert", "watching films", "smoking", and others related to lifestyle. Process crawling is done in Google Collaboratory. The data obtained is downloaded in .csv format.

2. **Sorting and labeling**

   The data that has been downloaded will be sorted so that you get data that does not deviate from the topic to be discussed. Next is labeling raw data for each tweet in the form

of sentences with positive, negative, and neutral polarities.

3. Pre-processing Data

Level pre-processing data consists of case-folding, tokenizing, stopword removal, and lemmatizing. This process is carried out using the Python programming language, software Google Collaboratory, library pandas, Sastrawi, and nltk.

The main goal of preprocessing data is tidying up messy initial data to be more structured so that it can be processed better. The following are the stages that are usually carried out in preprocessing data for sentiment analysis.

a. Case-Folding

Steps Case-folding is an algorithm that converts all characters in data to lowercase and removes repeated letters in a word. This is done to standardize the use of uppercase and lowercase letters in the same word so that these words are not considered different. For example, the words "Chicken" and "chicken" which are the same will feel odd if they are grouped into two different word categories just because of the difference in letters (Denny & Spirling, 2018).

b. Tokenization

Level Tokenization is a step to parse data that is still in sentence form into parts of words called tokens or single-word units (Ananda & Pristyanto, 2021).

c. Stopword Removal

This stage aims to eliminate words without meaning or stopwords and retain relevant words. Example stopwords in Indonesian include words such as "di", "yang", "pada", "untuk", and others. Many libraries in Python provide groups of stop words to simplify the deletion process. As an example in Indonesian, Sastrawi is used as an one library that provides this feature.

d. Lemmatizing

Level Lemmatizing is the step where a word is changed into a basic word form by considering the context of the word.

4. Exploratory Data Analysis (EDA)

Includes a critical process of initial investigative testing on data to identify patterns, discover anomalies, test hypotheses, or check assumptions through summary statistics and graphical representations. Thus, EDA can distinguish data sets that can be expressed further beyond formal data modeling or hypothesis testing tasks. In this process, data will be formed frame and divided into each class. This process is carried out using the software Google Collaboratory.

5. Data Visualization using Word Cloud

This stage aims to make the visual data more attractive and easy to analyze. A Word Cloud is an image consisting of a collection of words, where the size of the word represents the occurrence or level of importance of the word. The more a word appears, the more often a word is mentioned in a text document. The initial step in this stage is to categorize the data with negative, neutral, and positive polarity. This process is carried out using the software Google Collaboratory.

6. Feature Extraction from Text

Before performing feature extraction, it will be carried out splitting or division of data. Data will be divided into two, namely data training and data testing. Training data is used to train the model that will be formed. Meanwhile, testing data is used to test the performance of previously formed models. Feature extraction is the process of changing original data into numerical form this process can be carried out while maintaining the information originating from the original data. This process is carried out using the software Google Collaboratory.

Term Frequency–Inverse Document Frequency (TF-IDF) is an algorithm for giving weight to words in a document. Weights are given by considering the significance of words based on their frequency of occurrence. TF-IDF is an algorithm that combines two formulas, namely TF and IDF.

1. Term Frequency (TF)

TF is a weighting algorithm that measures how often a word appears in a document.

1. $TF_{ij} = f_{ij}$,

TF_{ij}: the weight possessed by term (j) in the document (i),
f_{ij}: total appearance term (j) in the document (i).

2. *Inverse Document Frequency* (IDF)

IDF is a weighting algorithm that measures how many times a word appears in various documents. IDF is the counter to TF weighting which depends on how rarely the word appears in the document. IDF gives lower weights to words that appear frequently while giving higher weights to words that appear rarely (Qaiser & Ali, 2018).

$$IDF_j = \log\left(\frac{D}{df_j}\right),$$

IDF_j: IDF weight for the term (j),
D: the total number of documents in the dataset,
df_j : total number of documents containing the term (j), > 0.

After that, the second output from this algorithm will be multiplied to produce a weight value for the words in the document.

TF-IDF produces superior performance compared to other weighting algorithms such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) (Adebiyi et al., 2019), and Okapi BM25 (Kadhim, 2019).

7. Make models

a) Model NBC

The NBC multinomial model that has been formed will be trained on data training to form a suitable model. The model will then be evaluated using a library classification report.

The steps in the classification process using NBC are as follows (Wilianto et al., 2017).

a. Shaping Features

Features refer to keywords that act as parameters in each training data, such as in documents (tweets), which are then classified into predefined categories (positive, negative, or neutral).

b. Calculating Probability P(c_i)

Calculate the probability for each class as follows.

$$P(c_i) = \frac{nd(c_i)}{|D|},$$

$nd(c_i)$: Number of documents included in the class,
$|D|$: The amount of training data or comments available.

c. Determining the Probability $P(w_k|c_i)$

Calculation of the probability of word occurrence in each class is done in the following way.

$$P(w_k|c_i) = \frac{f(w_{ki},c_i)+1}{f(c_i)+|V|}$$

$f(w_{ki}, c_i)$: Frequency of occurrence of words in class,
(c_i): The number of all words in the class,
$|V|$: The number of words in the training data.

d. Calculating Vmap

Vmap is a formula used by NBC to estimate the probability of test data for each class based on the previous steps. The Vmap equation can be stated as follows.

$$\text{Vmap} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_i P(a_i|v_j)$$

e. Determines the Maximum Vmap

Based on the calculation of the Vmap value for positive and negative classification, the highest Vmap value was selected as the classification result. For example, if a document has a positive Vmap value that is greater than a negative Vmap value, then the document will be classified as a positive classification result.

b) Model SVM

The SVM model that has been formed will be trained on data training to form a suitable model. The model will then be evaluated using a library classification report.

The initial process in SVM is the identification support vector in every class. The support vector is an example of data from each class that has the closest distance to each other. In this research, the support vector consists of positive polarity data and negative polarity data. After support vector is successfully identified, the next step is to calculate the margin. Margin is the distance between the support vector which serves as a separator between the two classes. In SVM, there are two types of margin, namely low margin and high margin. The low margin is the distance between the

support vectors which are close to each other, temporary high margin is the distance between the support vector which is far away. SVM aims to find the margin largest which effectively separates the two classes (Lombu et al., 2022).

The hyperplane has a maximum margin distance (maximum-margin hyperplane) which is measured from the hyperplane to the nearest vector. Equation for the hyperplane which is located on support vector are as follows:

$$w.x + b = 0\ ,$$

by using a normal vector that is perpendicular to the hyperplane, as a variable, and as a bias term. If the given training data can be split linearly, then there may be two hyperplanes that are chosen in such a way that no points lie between them and maximize the distance between them hyperplane these are (Berry & Kogan, 2010)

$$margin = \frac{2}{\|\vec{w}\|}.$$

It is not always possible to separate data using linear equations because there are situations where data has dimensions that cannot be separated linearly. Therefore, the solution is to use the kernel function to add dimension to the hyperplane. Kernel function is a mathematical method that allows SVM to carry out classification in a higher dimensional space than the dimensions of the original data. For example, if data initially has one dimension, using kernels, SVM can perform classification in two or more dimensions. Table 1 shows several kernel functions that can be applied to improve classification (Phienthrakul et al., 2009).

Table 1: Kernel function in SVM classification

Kernel	Formula
Linear	$K(x,y) = x.y$
Polinomial	$K(x,y) = (x.y + c)^d$
Gaussian Radial Basis Function	$K(x,y) = \exp(-\gamma\|x-y\|^2)$
Sigmoid	$\boldsymbol{K(x,y) = \tanh(x.y + c)}$

8. Model testing

Model testing is carried out by predicting the data testing using the NBC and SVM models obtained in stage seven. This process is carried out using the software Google Collaboratory.

9. Benchmarking performance level results

Evaluate the performance level results of the SVM and NBC models obtained using test data based on accuracy, precision, recall, and F1-Score. The three evaluation metrics from the SVM and NBC algorithms were compared to determine which method is better for classifying lifestyle on Twitter social media. This process is carried out using the software Google Collaboratory.

Model evaluation is needed in the classification process to test the extent to which the calculations that have been carried out perform well. Testing using two parameters,

namely calculation of accuracy, precision, recall, and F1-score. In binary classification, the classification results can produce four different situations: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). Usually, these four situations are represented in the form confusion matrix as shown below.

Based on the results confusion matrix calculations can be carried out to evaluate the results of the model classification, including calculating the values of accuracy, precision, recall, and F1-Score.

1. **Accuracy**

   Accuracy, also known as degree of accuracy, measures the extent of datatweet that have been correctly identified in the test data, or the extent to which the system predictions approximate manual predictions. Calculating the accuracy value can be done using the following formula (Grandini et al., 2020).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

2. **Precision**

   Precision, also called correctness, is a measure of how sensitive the system is in making correct predictions, both in positive class and negative class data. Calculating the precision value can be done using the following formula (Grandini et al., 2020).

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}}.$$

3. **Recall**

   Recall refers to the system's ability to correctly identify information again in negative and positive classes (Azhari et al., 2021). In this case,recall measures the extent to which the system is successful in re-detecting relevant information. Calculating value recall can be done using the following formula (Grandini et al., 2020).

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}}.$$

4. **F1-Score**

   F1-Score is a metric that is often used to assess the performance of binary classification when one of its classes has a low frequency compared to other classes (Liptonet al., 2014). CountF1-Score can be done using the following formula (Grandini et al., 2020).

$$\text{F1-Score} = 2\left(\frac{precision \cdot recall}{precision + recall}\right).$$

10. Interpretation of results

Interpret the results of which classification is better.

11. Drawing conclusions

Draw conclusions about which classification results are better.

RESULTS AND DISCUSSION

1. Descript Data

Data collection was carried out on Twitter with the keywords "konser", "nonton film", "merokok", etc. from January to February 2023. Next, the data will be divided into 70% training data and data tests as much as 30% of the total data. The data that has been obtained

amounts to 11,000 pieces tweet and saved in .csv format.

2. Pre-processing data

1. Case-Folding

Table 2: Before and after case-folding

Before Case-Folding	After Case-Folding
Hi army OPEN EARLY BOOKSLOT AUGUSTD IN JAKARTA ONLY MEMBERSHIP DENGAN KETENTUAN ketentuan selengkapnya boleh di dm jika berminat Fee 300k tiket Tff Booked Slot 100k tiket mengurangi harga fee tiket nanti setelah full payment jastip konser AugustDTour fyp	hi army open early bookslot augustd in jakarta only membership dengan ketentuan ketentuan selengkapnya boleh di dm jika berminat fee 300k tiket tff booked slot 100k tiket mengurangi harga fee tiket nanti setelah full payment jastip konser augustdtour fyp
Aneh bgt ini konser aneh bgt masa project bisa beda2 per section	aneh bgt ini konser aneh bgt masa project bisa beda2 per section
Oh dia pengen permainan ini cepat selesai ternyata	oh dia pengen permainan ini cepat selesai ternyata

2. Stopword Removal

Table 3: Before and After Stopword removal

Before Stopword Removal	After Stopword Removal
hi army open early bookslot augustd in jakarta only membership dengan ketentuan ketentuan selengkapnya boleh di dm jika berminat fee 300k tiket tff booked slot 100k tiket mengurangi harga fee tiket nanti setelah full payment jastip konser augustdtour fyp	hi army open early bookslot augustd in jakarta only membership ketentuan ketentuan selengkapnya dm berminat fee 300k tiket tff booked slot 100k tiket mengurangi harga fee tiket full payment jastip konser augustdtour fyp
aneh bgt ini konser aneh bgt masa project bisa beda2 per section	aneh bgt konser aneh bgt project beda2 section
oh dia pengen permainan ini cepat selesai ternyata	oh pengen permainan cepat selesai

3. Lemmatizing

Table 4: Before and after lemmatizing

Before Lemmatizing	After Lemmatizing
hi army open early bookslot augustd in jakarta only membership ketentuan ketentuan selengkapnya dm berminat fee 300k tiket tff booked slot 100k tiket mengurangi harga fee tiket full payment jastip konser augustdtour fyp	hi army open early bookslot august in jakarta only membership ketentuan ketentuan lengkap dm minat fee 300k tiket tf booked slot 100k tiket kurang harga fee tiket full payment jastip konser augustdtour fyp
aneh bgt konser aneh bgt project beda2 section	aneh bgt konser aneh bgt project beda section
oh pengen permainan cepat selesai	oh pengen permainan cepat selesai

4. Tokenization

Table 5: Before and after tokenization

Before Tokenization	After Tokenization
hi army open early bookslot august in jakarta only membership ketentuan ketentuan lengkap dm minat fee 300k tiket tf booked slot 100k tiket kurang harga fee tiket full payment jastip konser augustdtour fyp (D1)	[‘hi’, ‘army’, ‘open’, ‘early’, ‘bookslot’, ‘augustd’, ‘in’, ‘jakarta’, ‘only’, ‘membership’, ‘ketentuan’, ‘ketentuan’, ‘lengkap’, ‘dm’, ‘minat’, ‘fee’, ‘300k’, ‘tiket’, ‘tf’, ‘booked’, ‘slot’, ‘100k’, ‘tiket’, ‘kurang’, ‘harga’, ‘fee’, ‘tiket’, ‘full’, ‘payment’, ‘jastip’, ‘konser’, ‘augustdtour’, ‘fyp’]
aneh bgt konser aneh bgt project beda section (D2)	[‘aneh’, ‘bgt’, ‘konser’, ‘aneh’, ‘bgt’, ‘project’, ‘beda’, ‘section’]
oh pengen permainan cepat selesai (D3)	[‘oh’, ‘pengen’, ‘permainan’, ‘cepat’, ‘selesai’]

3. Data Visualization using Word Cloud

(a). 

(b). 

(c) 

Figure 2. (a) Visualization of positive sentiment data, (b) visualization of negative sentiment data, and (c) visualization of neutral sentiment data which is used to show the relationship between the number of words and the increase in words that are frequently used and related to lifestyle topics.

4. TF-IDF Weighting

Table 6: Example of manual TF-IDF calculation

Term	TF-IDF		
	D1	D2	D3
ketentuan	0.3522	0	0
kurang	0.4771	0	0
jakarta	0.4771	0	0
tiket	0	0	0
beda	0	0.4771	0
pengen	0	0	0.4771
lengkap	0.4771	0	0
minat	0.4771	0	0
harga	0.4771	0	0
hi	0.4771	0	0
army	0.4771	0	0
aneh	0	0.3522	0

oh	0	0	0.4771
membership	0.4771	0	0
konser	0.1761	0.1761	0
bgt	0	0.3522	0
project	0	0.4771	0
permainan	0	0	0.4771
cepat	0	0	0.4771
…	…	…	…
selesai	0	0	0.4771

5. Division of Training Data and Test Data

The process of dividing training data and test data can be done in various ways. In this study, training data and test data were divided by a percentage of 70%:30%. Modelclassifier trained using 70% of the data owned and tested using 30% of the data to test the performance of the model in question.

6. Support Vector Machine (SVM) Algorithm Modeling

The next process is data classification using SVM. Classification usessoftware Google Collaboratory and library sklearn svm module. SVM uses SVC, namely the condition where the hyperlane is linear.

Table 7: The result of SVM evaluation

Classifier	Accuracy	Presisi	Recall	F1-Score
SVM	0.61	0.61	0.61	0.61

Accuracy indicates the accuracy of the classifier. The accuracy value is one of the assessment parameters of the method that has been used. The accuracy value is obtained from the amount of data that has been successfully classified correctly according to its sentiment class from the entire amount of data that has been classified. A high accuracy value is obtained when a lot of data is successfully classified correctly according to the sentiment class. The higher the accuracy value, the higher the precision value will be, and vice versa, because the precision value is the number of positive classes that are correctly classified as positive class divided by the total data that is classified as positive class (Buntoro, 2016).Recall shows the ratio of positive predictions to total positive predictions. F1-Score shows the level of accuracy of the model dataset.

Based on the results of classification evaluation using SVM, the evaluation values using linear kernels produce the same values between accuracy, precision,recall, and F1-Score, namely 61%. The results of manual data labeling using the SVM method produced an overall accuracy rate of 61%.

7. Naïve Bayes Classifier (NBC) Algorithm Modeling

The next classification method is to use the NBC algorithm, namely multinomial NBC. The NBC classification process is carried out in software Google Collaboratory and library

sklearn with MultinomialNM module.Syntax NBC classification is attached in the Appendix. Before entering the classification process, a process is carried out.

1. Process Learning NBC

As an algorithm supervised learning, in the NBC classification process it is necessary to carry out a process of learning. An example of manual classification using six data points taken randomly from research data is shown in Table 7.

Table 7: Training data and sentiment

Training Data	Sentiment
[‘konser’, ‘osaka’, ‘ya’]	positif
[‘dah’, ‘kek’, ‘senter’, ‘konser’]	positif
[‘tiket’, ‘konser’, ‘raisaku’, ‘kejual’, ‘sedih’, ‘gajadi’, ‘nonton’]	negatif
[‘dilarang’, ‘merokok’]	netral
[‘denger’, ‘denger’, ‘habis’, ‘jual’, ‘tiket’, ‘konser’, ‘8jt’]	netral
[‘permainan’, ‘bola’, ‘bunder’, ‘dipimpin’, ‘pemain’, ‘blunder’]	negatif

Table 8: Example of the number of occurrences of keywords as features

Data	Feature (occurrence)	Sentiment
D1	konser (1), osaka (1), ya (1)	positif
D2	dah (1), kek (1), senter (1), konser (1)	positif
D3	tiket (1), konser (1), raisaku (1), kejual (1), sedih (1), gajadi(1), nonton (1)	negatif
D4	dilarang (1), merokok (1)	netral
D5	denger (2), habis (1), jual (1), tiket (1), konser (1), 8jt (1)	netral
D6	permainan (1), bola (1), bunder (1), dipimpin (1), pemain (1), blunder (1)	negatif

Table 8 contains six data used as training data, for example D1 to D6. After forming features with the appearance of words from the training data, the process continues by calculating the probability of each class as follows using equation 2.11. The class probabilities of the training data are shown in Table 9.

Table 9: Example of training data class probabilities.

Kelas sentimen (c_i)	Data						$nd(c_i)$	$P(c_i)$
	D1	D2	D3	D4	D5	D6		
positif (c_1)	1	1	0	0	0	0	2	$\frac{2}{6}$
negatif (c_2)	0	0	1	0	0	1	2	$\frac{2}{6}$
netral (c_3)	0	0	0	1	1	0	2	$\frac{2}{6}$

Calculation of the class probability of the training data or prior probability in the model is done by dividing the amount of data with positive, negative or neutral sentiment by the total amount of training data used. Table 8 shows two of the six data with positive sentiment, namely D1 and D2, so the positive class probability or positive prior probability is 0.3. The same is true for negative and neutral class probabilities or negative and neutral prior probabilities, namely D3 and D6 and D4 and D5. The next step is to calculate the probability

of each feature in each class using the equations discussed above. An example of manual calculation of feature probabilities for the first word in the positive, negative and neutral classes is as follows.

$$P(\text{konser}|\text{positif}) = \frac{2+1}{8+30} = 0.0789$$
$$P(\text{konser}|\text{negatif}) = \frac{1+1}{12+30} = 0.0476$$
$$P(\text{konser}|\text{netral}) = \frac{1+1}{10+30} = 0.05$$

The positive class probability calculation is carried out using the equation discussed above where the value two is the number of occurrences of the "concert" feature in the data with the positive class, then the number eight is the number of all features in the positive class, and the number 30 is the number of all features in the training data. The probability of each feature is shown in Table 10.

Table 10: Example of probability for each feature in the data.

Data	**Kelas Sentimen**		
	Positif (c_1)	**Negatif (c_2)**	**Netral (c_3)**
konser	0.0789	0.0476	0.0500
osaka	0.0526	0.0238	0.0250
ya	0.0526	0.0238	0.0250
dah	0.0526	0.0238	0.0250
kek	0.0526	0.0238	0.0250
senter	0.0263	0.0526	0.0250
tiket	0.0263	0.0476	0.0500
raisaku	0.0263	0.0476	0.0250
kejual	0.0263	0.0476	0.0250
sedih	0.0263	0.0476	0.0250
gajadi	0.0263	0.0476	0.0250
nonton	0.0263	0.0476	0.0250
dilarang	0.0263	0.0238	0.0500
merokok	0.0263	0.0238	0.0500
Data	**Kelas Sentimen**		
	Positif (c_1)	**Negatif (c_2)**	**Netral (c_3)**
denger	0.0263	0.0238	0.0750
habis	0.0263	0.0238	0.0500
jual	0.0263	0.0238	0.0500
8jt	0.0263	0.0238	0.0500
permainan	0.0263	0.0500	0.0250
bola	0.0263	0.0500	0.0250
bunder	0.0263	0.0500	0.0250
dipimpin	0.0263	0.0500	0.0250
pemain	0.0263	0.0500	0.0250
blunder	0.0263	0.0500	0.0250

2. NBC Classification Process

The next step in NBC classification is to find the highest probability from a test data. For example, take one test data that has been passed pre-processing which contains the words "day", "this", "concert", and "osaka". The calculation process uses the

equation discussed above using the feature probability values in Table 8. Positive probability and consecutively negative too as follows.

$$P(c_1)\prod p(a_i|c_1) = P(c_1)p(\text{"konser"}|c_1)p(\text{"osaka"}|c_1)p(\text{"ya"}|c_1)$$
$$P(c_1)\prod p(a_i|c_1) = (0.3)(0.079)(0.053)(0.053)$$
$$P(c_1)\prod p(a_i|c_1) = 6.65 \times 10^{-5}$$
$$P(c_2)\prod p(a_i|c_2) = P(c_2)p(\text{"konser"}|c_2)p(\text{"osaka"}|c_2)p(\text{"ya"}|c_2)$$
$$P(c_2)\prod p(a_i|c_2) = (0.3)(0.048)(0.024)(0.024)$$
$$P(c_2)\prod p(a_i|c_2) = 8.29 \times 10^{-6}$$
$$P(c_3)\prod p(a_i|c_3) = P(c_3)p(\text{"konser"}|c_3)p(\text{"osaka"}|c_3)p(\text{"ya"}|c_3)$$
$$P(c_3)\prod p(a_i|c_3) = (0.3)(0.05)(0.025)(0.025)$$
$$P(c_3)\prod p(a_i|c_3) = 9.38 \times 10^{-6}$$
$$\text{Vmap} = \underset{c_j \in C}{argmax}\, P(c_j)\prod_i P(a_i|c_j)$$

Based on the three probability calculations above, it is known that the classification probability for positive sentiment is greater than for negative and neutral sentiment. Therefore, the data tested is classified as data with a positive sentiment.

3. Comparison of Classification Results with SVM and NBC

Based on the classification that has been carried out, the next step is to compare matric Best evaluation of SVM and NBC algorithms. A comparison of evaluation metrics is shown in Table 11.

Table 11: Comparison of evaluation of SVM and NBC algorithms.

Algorithm	Accuracy	Presisi	Recall	F1-Score
SVM	0.61	0.61	0.61	0.61
NBC	0.59	0.63	0.58	0.57

Comparison of classification results is prioritized by comparingF1-Score because the data used in classification is not balanced between data with positive, negative and neutral polarities. Based on the comparison results of accuracy, precision, recall, and F1-Score obtained metric The highest evaluation was achieved using the SVM method using a linear kernel which had the same value, namely 61%.

CONCLUSION

Based on research that has been carried out, people's reactions to lifestyle give more positive sentiments with a total of 4.134 positive sentiments, negative sentiments with a total of 3.646, and neutral with a total of 3.220. The results of lifestyle sentiment classification using SVM with a linear kernel produce a classification model that is quite good with a value F1-

Score 61%. Lifestyle sentiment analysis using NBC produces a fairly good classification model with values F1-Score 57%. Based on a comparison of evaluation metrics between SVM and NBC, it shows that the SVM algorithm produces the best sentiment analysis with values of accuracy, precision, recall, and value F1-Score the same, namely 61%.

REFERENCES

Adebiyi, A. A., Ogunleye, O. M., Adebiyi, M., & Okesola, J. O. (2019). A Comparative Analysis of TF-IDF, LSI and LDA in Semantic Information Retrieval Approach for Paper-Reviewer Assignment. *Journal of Engineering and Applied Sciences*, *14*(10), 3378–3382.

Ahmad, A., & Gata, W. (2022). Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Metaverse dengan Algoritma Support Vector Machine. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *6*(4), 548–555.

Amelia, R., Darmansah, D., Prastiwi, N. S., & Purbaya, M. E. (2022). Impementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Masyarakat Indonesia Mengenai Drama Korea Pada Twitter. *JURIKOM (Jurnal Riset Komputer)*, *9*(2), 338.

Ananda, F. D., & Pristyanto, Y. (2021). Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(2), 407–416.

Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. *ACM International Conference Proceeding Series*, *May*.

Artikel, I. (2006). Lifestyle. *Textile View Magazine*, *01*(73), 293–301.

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 640.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. *Studies in Computational Intelligence*, *639*, 313–340.

Berry, M. W., & Kogan, J. (2010). *Text mining: applications and theory*. John Wiley & Sons.

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, *26*(2), 168–189.

Dixon, S. (2022). *Countries with most Twitter users 2022*. Statista.

Fitri, E. (2020). Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. *Jurnal Transformatika*, *18*(1), 71.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17.

Kadhim, A. I. (2019). Term Weighting for Feature Extraction on Twitter: A Comparison between BM25 and TF-IDF. *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*, 124–128.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Lombu, A. S., Hidayat, S., & Hidayatullah, A. F. (2022). Pemodelan Klasifikasi Gaji Menggunakan Support Vector Machine. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 363–370.

Naufal, M. F., Arifin, T., & Wirjawan, H. (2023). Analisis Perbandingan Tingkat Performa Algoritma SVM, Random Forest, dan Naïve Bayes untuk Klasifikasi Cyberbullying pada Media Sosial. *Jurasik (Jurnal Riset Sistem …*, *8*(1), 82–90.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends*

in Information Retrieval, *2*(1–2), 1–135.

Phienthrakul, T., Kijsirikul, B., Takamura, H., & Okumura, M. (2009). Sentiment classification with support vector machines and multiple kernel functions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5864 LNCS*(PART 2), 583–592.

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1), 25–29.

Wilianto, L., Pudjiantoro, T. H., & Umbara, F. R. (2017). Analisis Sentimen Terhadap Tempat Wisata Dari Komentar Pengunjung dengan Menggunakan Metode Naïve Bayes Classifier Studi Kasus Jawa Barat. *Prosiding SNATIF*, 439–448.