Received December 16th, 2024; accepted December 23rd, 2024. Date of publication December 30th, 2024
Digital Object Identifier: https://doi.org/10/25047/jtit.v11i2.439

Sentiment Analysis of the Use of Makeup Products Using the Support Vector Machine Method

KHAIRUNNISA[1], SRIANI[2]

[1,2] Computer Science, Faculty of Science and Technology, State Islamic University of North Sumatera, Indonesia

CORESPONDING AUTHOR: KHAIRUNNISA (email : nisalubis1510@gmail.com)

ABSTRACT Many beauty products have emerged from various brands by providing attractive offers for women who are their main targets. Product reviews can help consumers regarding the quality of using the product. However, the problem is, on the femaledaily.com website there is no distinction between negative, neutral, and positive reviews so that consumers must first read the review and it takes a lot of time and this problem really requires a classification process on the review into negative, neutral, and positive classes. This process cannot be done automatically, therefore sentiment analysis is needed. To find out the classification of positive, negative, and neutral sentiment on the product, the Support Vector Machine (SVM) method is used, the advantage of SVM in this case lies in its ability to handle high-dimensional datasets and still produce effective classification and SVM is also a good choice for sentiment analysis in the context of cosmetic product reviews. The classification results using the SVM method produce data into 3 classes, namely 510 positive reviews, 98 neutral, and 29 negative with an accuracy value of 77.97%, precision 78%, recall 100%, fi-score 88%

KEYWORDS: Product Review, Sentiment Analysis, Femaledaily, Support Vector Machine

I.INTRODUCTION

Nowadays, women, who are the primary market for cosmetics, consider them to be necessities [1]. Therefore, more and more makeup brands have sprung up both domestically and abroad. Based on research from SAC (Science Art Communication) Indonesia, Indonesia is currently a market for cosmetic products throughout 2018. The skincare market contributed US$2,022 million to the cosmetics market and US$5,502 million to body care [2].

Somethinc products are a makeup and skincare brand that is well-known among young people in Indonesia. Somethinc itself is a local beauty brand originating from Indonesia since 2019. Not inferior to outside beauty brands, Somethinc itself has a cushion consisting of many shades that are tailored to the skin tone of Indonesian women. But, with the many types of shades available, consumers must pay attention to which shades are suitable for their skin tone by looking at reviews from previous users.

Make-up is an art used to beautify yourself by using various cosmetics such as foundation, powder, lipstick, mascara, eyeliner, and many more types. Make-up is usually synonymous with women who want to look more attractive. With make-up, women can freely experiment and create according to their own desires. Usually, make-up is used to cover black spots, acne scars on the skin of the face, to make our faces look fresher.

Hooman Somethinc Cushion uses Breathable Technology which is anti-bacterial & does not easily oxidize even if worn all day. Hooman Somethinc Cushion has Medium to High Coverage that is able to disguise pores and fine lines in one tap, and produces a longlasting Skin Matte Finish. This cushion contains SPF35 PA++++ [3]

On the Femaledaily.com website, there are many stores that sell cushion makeup with various brands and shades. Each brand has a cushion with its own review according to the experience of consumers who have used it. But unfortunately, on this website the positive sentiment and negative sentiment of each type are united so that new consumers have to read all the existing product reviews first so it takes a long time to see the review as a whole.

Technically, in providing a review of a product or service that has been used, namely by users providing an assessment based on the level of user satisfaction. The purpose of the review is to

provide information to buyers, provide confidence in the product or service to be used and provide an evaluation for sellers [4]

The goal of opinion mining, often referred to as sentiment analysis, is to ascertain if a body of written material (documents, sentences, paragraphs, etc.) tends to be positive, negative, or neutral in terms of polarity [5]. Text is typically classified into three sentiment groups using sentiment analysis: positive, negative, and neutral sentiment, a review can be considered in the positive sentiment class if it contains statements of praise, expressions of gratitude or positive reviews for a product [6].

Based on the outcomes of the training process, Support Vector Machines (SVM) can forecast classes using machine learning, also known as supervised learning. A pattern that will subsequently be utilized in the labeling process is produced by training with numerical input data and feature extraction findings [7]. SVM works by finding the best hyperline to divide classes in the input space, from research [8] said that the SVM algorithm has a better AUC value than naive bayes.

In the classification model, Support Vector Machine has a stronger and clearer mathematical concept. Support Vector Machine is used to find the optimal hyperplane by maximizing the distance between classes [9]. Support Vector Machine uses 2 points (vectors) which then the two points will form a boundary line (the boundary side if 3 dimensions or more) the boundary line formed from these two vectors is called a hyperplane. The concept of SVM can be summed up as identifying the optimal hyperplane in the input space that acts as a separator between two classes. This helps distinguish between positive (+1) and negative (-1) data classes. Positive (+1) data is shown in yellow in Figure 2.1, while negative (-1) data is shown in red. SVM maximizes the distance between classes in an attempt to find a separation function, or hyperplane [10].

The process of classifying a data object involves evaluating the data object to assign it to one of several accessible classes. In classification, the two main tasks to be performed are:

a. Building a prototype model that will be stored in memory. Classification algorithms are used to analyze each record in the training data according to its attribute values to create the model.

b. Other data objects are recognized, classified, and predicted using the model, thus allowing one to determine which class of data objects belong to the stored model. At this stage, the data is checked to ensure the accuracy level of the final model.

The model can be used to categorize new data records that have not been evaluated before if the accuracy rate obtained matches the given value. By examining the training data, a classification algorithm will create a classification model. Forming a function or mapping y = f(x), where x is the record whose class is to be estimated and y is the class with the expected result, this is another way to conceptualize the learning stage [11].

The classification model can be assessed using test data that has a specific value but is not used for training data. Model categorization is the process of creating a variety of data representations using target prediction outputs or data classes. A classification that yields two class outputs is known as binary classification {positive, negative} is used to describe the two classes. Accuracy is a metric used to assess categorization models. Accuracy is the sum of the ratio of correct predictions. This evaluation serves to measure the accuracy value using K-fold cross validation and Yeh 2020 [10].

Based on this background, the author will design a system that can automatically help consumers to be able to shorten the time to see reviews of the use of these products on the female daily website using the SVM method. This method is taken because it can produce optimal results in classification and also several similar studies also use this method.

In research conducted by Putri with the title Sentiment Analysis on Beauty Products from Female Daily Reviews Using Information Gain and SVM Classifer, the result is that the analysis on beauty products is to use the Preprocessing process (Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stemming (without Stopword)), and also by using TF-IDF feature extraction, Information Gain feature selection using a threshold limit of 0.01, and using the SVM classification process which produces an accuracy value of 85.89% [12].

Previously in 2023 there was a previous study entitled “Sentiment Analysis on Cushion Products on the Female Daily Website Using the SVM Method” [13] which has similarities with this research because it uses the SVM method but, there are also differences between the two studies which in previous studies only focused on the pixy cushion brand and only produced two sentiments, namely positive and negative sentiments. Meanwhile, this research focuses on the somethinc hooman cushion brand and will produce three sentiments, namely positive, neutral, and negative.

II.METHOD

A. Support Vector Machine

In this research, sentiment analysis is carried out using the SVM method to find out the reviews of makeup products on the femaledaily.com website to make it easier for new consumers to see which products with shades that match their desires. The framework of this research is shown in Figure 1 below:

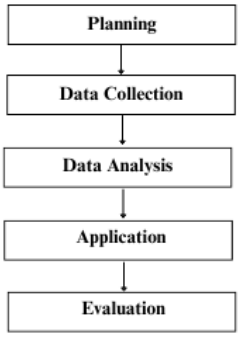

FIGURE 1. Research Outline

This research process begins with planning, namely determining the topic to be discussed. The topic of this research is sentiment analysis of the use of somethinc hooman makeup products using the support vector machine (SVM) method. Furthermore, automatic collection of review data using a web scrapper is carried out, totaling approximately 587 data starting from 2021 to 2024.

In this research, data analysis is carried out so that it can be carried out properly. At this stage there are needs for what is used in research, product review data which is the main focus. The SVM method is used to perform the sentiment analysis stage. Training and testing data is the first input, and is handled in the preprocessing step before being classified. The classification method will use data that already contains labels. This procedure is followed to create a model that will be applied to categorize new data.

The application of this research begins with retrieving data from the femaledaily.com web regarding consumer reviews about what will be classified, then calculating the weight of the word to get a sentiment classification using the SVM method then the dataset obtained will be classified using confusion matrrix so that it becomes three classes, namely negative class, positive class, and neutral class.

Using the confusion matrix and SVM library, the classified training and testing models were used in this research test to determine the accuracy value in predicting the sentiment of Somethinc Hooman makeup product reviews. Until a matrix representing the actual class and the predicted class is generated. Output training model with new data that has never been done before.

III.RESULT AND DISCUSSION

This research will examine sentiment data regarding somenthinc hooman makeup product reviews. Data collection is the first step in the research process. By using python as a data scraper, the process of collecting and consumer opinions about somethinc hooman makeup product reviews on the femaledaily website.

The data that will be taken is consumer reviews on somethinc hooman products and taken to do the data scrapping process. Reviews from January 01, 2022 to July 18, 2024 are the data to be collected. The output of the scrapping process will be saved with the existence of a.csv and then processed by the python programmer.

The data is grouped into three label categories, namely positive, neutral, and negative. This labeling process is done to facilitate sentiment analysis and help understand the perceptions or emotions contained in the data. A visualization of this label division can be seen in the figure below, which illustrates the distribution of data based on its category. These labels serve as the basis for further analysis steps, such as model training or sentiment pattern evaluation.

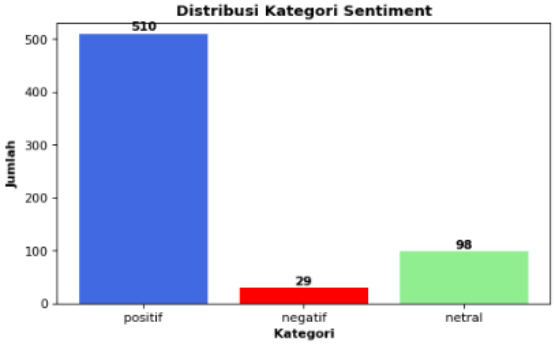

FIGURE 2. Sentiment Data Distribution

B. Pre-processing Data

Preprocessing is necessary before scraped data can be examined as best it can. The goal of this procedure is to remove noise, correct discrepancies, and organize and clean up unstructured data. Data that has undergone preprocessing will therefore be more prepared for analysis, leading to more accurate and trustworthy findings.

The scraping results in data with five columns, namely web_scraper_order, web_scraper_start_url, date, name, and reviews. However, for sentiment analysis purposes, only the `review` column will be used. Therefore, a data structure cleaning step is required to filter out irrelevant information and ensure only the `review` column is retained for the further analysis process.

As an example of manual calculation in this study, 4 training data are provided as samples, namely:

TABLE 1. Data Training

No	Reviews	Pre-Processing Result
1	beli shade yang paling terang buat undertone neutral, tapi tetep kurang terang buat kulitku, cuman masih bisa diatur sih gak sampe kegelapan banget semi matte dan cuman perlu di set dikit pake bedak supaya gak transfer	"neutral", "semi", "matte", "bedak"

No	Reviews	Pre-Processing Result
2	cushion satu ini ga diraguin lagi deh. buat yang aku punya kulit muka oily dan ada beberapa darkspot langsung tercover sempurna. dipakai buat aktifitas luar juga oke, bedannya tetep on fire lah make up dimuka kita dan ringan banget berasa ga pake cushion	"darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "beda"
3	one and only my lovely cushion. cushion yang ringan dan coverage nya mantul. yang aku suka dari cushion ini dia finish nya ga lebay gitu, jadi kayak second skin aja. trs kalo dipake makin lama makin cakep, minim oksidasi bahkan hampir mendekati gaada oksidasinya.	"ragu", "darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "beda", "ringan", "rasa"
4	setiap pakai ini selalu jerawatan entah kenapa bisa breakout massal. buat performanya oke oke aja cuman sayangnya di kulitku yg sensitif ini ga cocok banget. warnanya di aku masih sedikit kegelapan padahal nina shade paling light di neutral. harusnya ada yg lebih terang dari nina sih	"jerawat", "breakout", "performa", "sensitif", "cocok", "gelap", "neutral", "lebih", "terang"

C. TF-IDF

Weighting utilizing TF-IDF (Term Frequency-Inverse Document Frequency) comes after the labeling and sentiment cleaning phases. Term Frequency (TF), which gauges how frequently a word occurs in a document, and Inverse Document Frequency (IDF), which lowers the weight of terms that occur frequently in numerous documents, are the two factors used to determine each word's weight at this point. By assigning larger weights to words that are rarely used in other documents, the results of the TF and IDF computations are multiplied to estimate the word's value in the context of the document.

TABLE 2. Label of Data Training

No	Data Training	Label
1	["neutral", "semi", "matte", "bedak"]	Negative
2	["darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "beda"]	Neutral
3	["ragu", "darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "beda", "ringan", "rasa"]	Positve
4	["jerawat", "breakout", "performa", "sensitif", "cocok", "gelap", "neutral", "lebih", "terang"]	Positve

TABLE 3. DF Value

Term	TF				DF
	D1	D2	D3	D4	
neutral	1	0	0	1	2
semi	1	0	0	0	1
matte	1	0	0	0	1
bedak	1	0	0	0	1
darkspot	0	1	1	0	2
cover	0	1	1	0	2
sempurna	0	1	1	0	2
pakai	0	1	1	0	2
aktifitas	0	1	1	0	2
luar	0	1	1	0	2
beda	0	1	1	0	2
ragu	0	0	1	0	1
ringan	0	0	1	0	1
rasa	0	0	1	0	1
jerawat	0	0	0	1	1
breakout	0	0	0	1	1
performa	0	0	0	1	1
sensitif	0	0	0	1	1
cocok	0	0	0	1	1
gelap	0	0	0	1	1
lebih	0	0	0	1	1
terang	0	0	0	1	1

After the TF (Term Frequency) value is calculated, the next step is to determine the IDF (Inverse Document Frequency) value. The following is the formula used to calculate the IDF value for each word.

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

with:
IDF: Inverse Document Frequency
D : word frequency in D

df : many documents containing the search word

The following is a sample in applying the formula to the first and second data:

$$IDF\ neutral = \log\left(\frac{D+1}{df+1}\right) + 1$$

$$= \log\left(\frac{4+1}{2+1}\right) + 1$$

$$= \log\left(\frac{5}{3}\right) + 1$$

= 1,2218

$$IDF\ semi = \log\left(\frac{D+1}{df+1}\right) + 1$$

$$= \log\left(\frac{4+1}{1+1}\right) + 1$$

$$= \log\left(\frac{5}{2}\right) + 1$$

= 1,3979

After obtaining the TF and IDF values, the next step is to calculate the TF-IDF value. This calculation is done using the following equation.

$$W = TF \times IDF \quad (2)$$

with:

W : the weight of the dth document against the tth word

TF : number of words in the searched document

IDF : Inverse Document Frequency

Here is a sample of applying the formula to the first data:

$$Wneutral = TF \times IDF$$

$$= 2 \times 1{,}2218$$

= 2,4437

$$Wsemi = TF \times IDF$$

$$= 1 \times 1{,}3979$$

= 1,3979

For further results in Table 4 using the same calculations as above. The blue part of the table is the result of the calculations done above.

TABLE 4. TF-IDF

Term	TF-IDF			
	D1	D2	D3	D4
neutral	2,4437	0	0	2,4437
semi	1,3979	0	0	0
matte	1,3979	0	0	0
bedak	1,3979	0	0	0
darkspot	0	2,4437	2,4437	0
cover	0	2,4437	2,4437	0
sempurna	0	2,4437	2,4437	0
pakai	0	2,4437	2,4437	0
aktifitas	0	2,4437	2,4437	0
luar	0	2,4437	2,4437	0
beda	0	2,4437	2,4437	0
ragu	0	0	1,39794	0
ringan	0	0	1,39794	0
rasa	0	0	1,39794	0
jerawat	0	0	0	1,39794
breakout	0	0	0	1,39794
performa	0	0	0	1,39794
sensitif	0	0	0	1,39794
cocok	0	0	0	1,39794
gelap	0	0	0	1,39794
lebih	0	0	0	1,39794
terang	0	0	0	1,39794

The next stage is the normalization of TF-IDF values to align the intervals in each data. This normalization is a fundamental step in data mining which aims to maintain the consistency of each record in the dataset.

The data was normalized using the following equation.

$$TF_{norm}(t,d) = \frac{TF(t,d)}{\sqrt{\sum_i (TF(t,d))^2}} \quad (3)$$

With:

d : dth document

t : t-th word of the keyword

TF : the number of words in the document being searched

Samples in applying the formula include:

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

$$= \frac{2{,}4437}{\sqrt{(2{,}4437)^2 + (1{,}3979)^2 \ldots + (1{,}3979)^2}}$$

$$= \frac{2{,}4437}{\sqrt{11{,}08629}}$$

= 0,220425

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

$$= \frac{1{,}3979}{\sqrt{(2{,}4437)^2 + (1{,}3979)^2 \ldots + (1{,}3979)^2}}$$

$$= \frac{1{,}39799}{\sqrt{11{,}08629}}$$

= 0,126096

For further results in Table 5 using the same calculations as above. The blue part in the table is the result of the calculation done above. The outcomes of the data normalization computations are as follows.

TABLE 5. Data Normalization

D1	D2	D3	D4
0,220425	0	0	0,220425

D1	D2	D3	D4
0,126096	0	0	0
0,126096	0	0	0
0,126096	0	0	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0,220425	0,220425	0
0	0	0,126096	0
0	0	0,126096	0
0	0	0,126096	0
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096
0	0	0	0,126096

D. Classification Support Vector Machine

The Support Vector Machine (SVM) algorithm is then used to apply classification after the data has been appropriately cleaned and arranged. Separating the dataset into training and test data is the first step in this procedure. The three classes—positive, neutral, and negative—are taught their patterns and traits using the training data. In the meantime, the test data is used to assess how well the model performs accurate categorization.

The training and test data used in this investigation are listed below. While test data is used to assess how well the model performs in making predictions or classifications on previously unseen data, training data is used to teach the model to recognize specific patterns or traits. The goal of this data split is to guarantee that the final model has strong generalization skills.

TABLE 6. Data Training of SVM

No	Data Training	Label
1	["neutral", "semi", "matte", "bedak"]	Negative
2	["darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "beda"]	Neutral
3	["ragu", "darkspot", "cover", "sempurna", "pakai", "aktifitas", "luar", "bedan", "ringan", "rasa"]	Positive
4	["jerawat", "breakout", "performa", "sensitif", "cocok", "gelap","neutral", "lebih", "terang"]	Positive

TABLE 7. Data Testing of SVM

Data Testing
["hydrate", "matte", "kering", "cocok",]

A linear kernel is the kind of kernel utilized in the classification process. A linear kernel is better suited for efficiently separating data in feature space since it is chosen based on the linear properties of the data. The linear kernel calculates the relationship between data by projecting it into a high-dimensional space without complex transformations. The following equation is used to calculate the value of the linear kernel, which is generally expressed as follows:

$$K(x, y) = x.y \quad (4)$$

Here is an example of data representation on 4 pieces of data.

TABLE 8. Sample Data Representation

	x1	x2	x3	x4
y1	K(x1,y1)	K(x2,y1)	K(x3,y1)	K(x4,y1)
y2	K(x1,y2)	K(x2,y2)	K(x3,y2)	K(x4,y2)
y3	K(x1,y3)	K(x2,y3)	K(x3,y3)	K(x4,y3)
y4	K(x1,y4)	K(x2,y4)	K(x3,y4)	K(x4,y4)

The results of the linear kernel calculation based on the sample data can be seen in the table below.

Example calculation for column 1 row 1.

K (x,y) = x * y

K (x,y) = (t1d1 * t1d1 + t1d2 * t1d2 + t1d3 * t1d3 + t1d4 * t1d4 + t1d5 * t1d5)

K (x,y) = (0,220425 * 0,220425 + 0*0 + 0*0 + 0,220425*0,220425)

D1,1 = **0.0972**

After the kernel value is obtained, the next step is to calculate the Hessian matrix. The following is an explanation of each parameter used in the calculation of the Hessian matrix [14].

The calculation is based on the following equation, which is designed to calculate the elements in the Hessian matrix. This approach aims to ensure

the calculation results are accurate and can be used in the next optimization stage.

The equations used are:

$$H_{ij} = y_i y_j (K(x_i, x_j) + \lambda^2 \quad (5)$$

With:
i,j = 1, 2, ..., n
K(xi, xj) = kernel value (normalization)
yi = i-th data class
yj = j-th data class

Example of calculating the value of the Positive Hessian matrix in column 1 row 1

H1,1 = $y_i y_j (K(x_i, x_j) + \lambda^2$

= -1 * -1 (0,0972) + 0.52

= 0,3472

After completing the calculation of the Hessian matrix value, the next step is to perform sequential training calculations. This process is carried out using predetermined equations to optimize the model parameters gradually [15]. The sequential training approach aims to update the model parameters iteratively, resulting in an optimal solution that meets the desired criteria. The equations used in this stage are as follows.

$$\nabla wF(w)=Hw+b \quad (6)$$

With:
∇wF(w) = gradient of the objective function F(w)
Hw = Hessian matrix
b = bias

To start the iteration calculation process in optimization, we assume the initial value of the weight vector (w) is 0.1, the bias (b) is 0.5, and the learning error rate is set at 0.01. These initial parameters will be used as the basis for updates in each iteration step until it reaches convergence or the desired minimum error.

∇wF(w) kolom 1 positive = Hw+b

∇wF(w) =(0,3472 +0,2778+….+0,2778+0,2778) × 0,1 + 0,5

∇wF(w) =(0,54582+0,5) = **1,04582**

Next, we will calculate the new weights and biases using the following equations

$$wnew\ kolom\ 1 = wold - \eta \nabla wF(w) \quad (7)$$

wnew kolom 1=0.1−(0.01 x 1,045)
wnew kolom 1= **0,0895**

$$bnew = bold - \eta \nabla bF(b) \quad (8)$$

bnew= 0,5 – (0,01*(0,1+0,1+….+0,1+0,1)) = **0,478**

Thus, after the calculation is done, the new bias value obtained is 0.478. This value illustrates the results of the bias update based on the iteration process that uses the gradient to minimize the error function.

Next, the process will continue to calculate the value of iteration 1 to further update the model parameters. here is the value of iteration 1 in column 1 row 1.

∇wF(w) kolom 1 baris 1 positive = Hw+b

∇wF(w) =(0,3472 +0,2778+….+0,2778+0,2778) × 0,08954 + 0,478

∇wF(w) =(0,4873+0,478) = 0,9753

For the next results in Table 9 using the same calculations as above. The blue-colored part in the table is the result of the calculation done above. Here are the results of the new weight calculations carried out.

TABLE 9. Iterasi I

No	Positive	Negative	Neutral
1	0,965349	0,964765	1,004837
2	0,975825	0,975241	0,97559
3	0,975825	0,975241	0,97559
4	0,975825	0,975241	0,97559
5	1,022083	1,036307	1,021861
6	1,022083	1,036307	1,021861
7	1,022083	1,036307	1,021861
8	1,022083	1,036307	1,021861
9	1,022083	1,036307	1,021861
10	1,022083	1,036307	1,021861
11	1,022083	1,036307	1,021861
12	0,956046	0,990012	0,955824
13	0,956046	0,990012	0,955824
14	0,956046	0,990012	0,955824
15	0,977953	0,977369	0,98269
16	0,977953	0,977369	0,98269
17	0,977953	0,977369	0,98269
18	0,977953	0,977369	0,98269
19	0,977953	0,977369	0,98269
20	0,977953	0,977369	0,98269
21	0,977953	0,977369	0,98269
22	0,977953	0,977369	0,98269

The iteration process was stopped at iteration 1 because the model had reached the optimal condition. This indicates that the predefined error tolerance has been met, so no additional iterations are required. This decision ensures computational time efficiency while maintaining the accuracy and quality of the resulting model.

E. Support Vector Machine

At this stage, test data will be used to measure the performance of the SVM algorithm that has been trained. This process aims to evaluate the model's ability to classify new data based on the weights and biases that have been obtained during training. The test results will provide an overview of the accuracy,

precision, and reliability of the model in solving classification problems.

For hydrate:

$IDF = \log\left(\frac{D+1}{df+1}\right) + 1$

$= \log\left(\frac{4+1}{1+1}\right) + 1$

$= \log\left(\frac{5}{2}\right) + 1$

= 1,3979

$W = TF \times IDF$

= 1 ×1,3979

= 1,3979

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

$= \frac{1,3979}{\sqrt{(2,4437)^2 + (1,3979)^2 \ldots + (1,3979)^2}}$

$= \frac{1,3979}{\sqrt{11,08629}}$

= 0,126096

The calculation is also performed on matte, dry, and matched data. So the results of these calculations are shown in Table 10 below:

TABLE 10. TF-IDF of Data Testing

Term	TF	DF	IDF	TF-IDF	Norm
hydrate	1	0	1,3979	1,3979	0,126096
matte	1	1	1,3979	1,3979	0,126096
kering	1	0	1,3979	1,3979	0,126096
cocok	1	1	1,3979	1,3979	0,126096

Next, we calculate the kernel of each test data with the previous training data. The following is an example of training data kernel calculation on column 1 row 1 test data where the results are in the blue column, and for the next results using the same calculation.

K (x,y) = x * y

K (x,y) = (t1d1 * t1d1 + t1d2 * t1d2 + t1d3 * t1d3 + t1d4 * t1d4)

K (x,y) = (0,126096 * 0,220425 + 0*0 + 0*0 + 0,126096 *0,220425)

D1,1 = **0,0556**

TABLE 11. Training Data Kernel Calculation Results Against Test Data

No	hydrate	matte	kering	cocok
1	0,0556	0,0556	0,0556	0,0556
2	0,0159	0,0159	0,0159	0,0159
3	0,0159	0,0159	0,0159	0,0159
4	0,0159	0,0159	0,0159	0,0159
5	0,0556	0,0556	0,0556	0,0556
6	0,0556	0,0556	0,0556	0,0556
7	0,0556	0,0556	0,0556	0,0556
8	0,0556	0,0556	0,0556	0,0556
9	0,0556	0,0556	0,0556	0,0556
10	0,0556	0,0556	0,0556	0,0556
11	0,0556	0,0556	0,0556	0,0556
12	0,0159	0,0159	0,0159	0,0159
13	0,0159	0,0159	0,0159	0,0159
14	0,0159	0,0159	0,0159	0,0159
15	0,0159	0,0159	0,0159	0,0159
16	0,0159	0,0159	0,0159	0,0159
17	0,0159	0,0159	0,0159	0,0159
18	0,0159	0,0159	0,0159	0,0159
19	0,0159	0,0159	0,0159	0,0159
20	0,0159	0,0159	0,0159	0,0159
21	0,0159	0,0159	0,0159	0,0159
22	0,0159	0,0159	0,0159	0,0159
∑	0,6673	0,6673	0,6673	0,6673

The sum of the weights from the test data is used to calculate the function value f(x) using the equation f(x)=w·x+b, where w is the weight vector, x is the input data vector, and b is the bias. This function describes the linear relationship between input and output, where w affects how much each data element contributes to the final result, and b adds a constant for result adjustment. After the value of f(x) is calculated, the next step is to determine the prediction or classification based on that value. And the resulting values are:

f(positive) = 0.71677
f(negative) = 0.71677
f(neutral) = 0.71648

In situations where the decision function values for the positive and negative classes have identical results, the SVM algorithm will determine the classification based on the predefined prioritization rules. Based on the rule, the data will be classified into the positive class as the final result. This shows that under the condition of the same function value, the class selection is done according to the algorithm's internal policy to ensure the decision remains consistent.

The following is the result of the confusion matrix that shows the performance of the model in classifying the test data. The confusion matrix provides an overview of the extent to which the predictions made by the model match the actual classes. This matrix is very important for evaluating the accuracy, precision, recall, and F1-score of the model, as well as providing more in-depth information about the strengths and weaknesses of the model in handling each class.

TABLE 12. Confussion Matrix

	precision	recall	f1-score	support
negative	1.00	0.00	0.00	4
neutral	1.00	0.00	0.00	22
positive	0.78	1.00	0.88	92
accuracy			0.78	118
macro avg	0.93	0.33	0.29	118
weighted avg	0.83	0.78	0.68	118

Accuracy indicates how well the model does in making predictions compared to the original data. Accuracy is calculated as the percentage of the number of correct predictions compared to the total predictions made. A high accuracy indicates that the model has a good ability to classify data, while a low accuracy indicates that the model still needs to be improved to produce more accurate predictions. The accuracy obtained from this system is 77.97 percent.

The results of this study show a higher accuracy value compared to research with the title Classification of Review Text Models on E-commerce Tokopedia Using SVM Algorithm conducted by Milal, where the results of the study got a percentage of 75% for the accuracy of this SVM classification model. The classification results also get the highest precision value in the positive tendency class with a value of 1.00, the highest recal - f1score - support value is in the same class, namely positive with a value of 0.76 - 0.57 - 106 respectively. And also the results of the classification on the website display the percentage of negative probability and positive probability [16].

IV.CONCLUSION

By using the SVM method, we can classify cosmetic product reviews with a fairly high accuracy, such as 77.97%. This shows that SVM can separate positive, negative, and neutral reviews quite well based on the features generated from the text data. In addition, this model also shows good ability in handling data with many feature variables, such as those found in text representation using the TF-IDF method. The advantage of SVM in this case lies in its ability to handle high-dimensional datasets and still produce effective classification. Overall, SVM is a good choice for sentiment analysis in the context of cosmetic product reviews. The data showed 510 positive, 98 neutral, and 29 negative reviews, with the majority of reviews having a positive sentiment. This division allows the SVM model to be trained to recognize sentiment patterns in the text, which is important for producing an accurate analysis of consumer perceptions of cosmetic products. By using SVM to classify cosmetic product reviews on Google Colab, the model can effectively separate positive, negative and neutral sentiments. With an accuracy of 77.97%, SVM proved to be good at text sentiment analysis, and the clear division of data helped the model learn text patterns better.

REFERENCE

[1] S. A. Ashari, M. W. A. Saputra, E. Larosa, and B. S. Rijal, "Analisis Sentimen pada Aplikasi Translate Google Menggunakan Metode SVM (Studi Kasus: Komentar Pada Playstore)," *jt*, vol. 21, no. 2, pp. 168–182, Dec. 2023, doi: 10.37031/jt.v21i2.412.

[2] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," no. 1, 2019.

[3] R. W. Pratiwi, S. F. H, D. Dairoh, D. I. Af'idah, Q. R. A, and A. G. F, "Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM)," *INISTA*, vol. 4, no. 1, pp. 40–46, Dec. 2021, doi: 10.20895/inista.v4i1.387.

[4] S. Ariqoh, M. A. Sunandar, and Y. Muhyidin, "ANALISIS SENTIMEN PADA PRODUK CUSHION DI WEBSITE FEMALE DAILY MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)," *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer*, vol. 2, no. 3, pp. 137–142, Aug. 2023, doi: 10.55123/storage.v2i3.2345.

[5] F. Ramadhanty, "PENGARUH LIVE SHOPPING TIKTOK @SKINTIFIC_INDONESIA TERHADAP KEPUTUSAN PEMBELIAN MAHASISWA PROGRAM STUDI ILMU KOMUNIKASI UNIVERSITAS FAJAR," 2023.

[6] P. B. Ansori and P. B. Ansori, "PENGARUH HARGA, LOKASI DAN KUALITAS PRODUK TERHADAP KEPUTUSAN PEMBELIAN KONSUMEN CV. ZAFIRA TEKNIK PEKANBARU," *Ekobis*, vol. 11, no. 1, pp. 11–19, Mar. 2020, doi: 10.36975/jeb.v11i1.253.

[7] A. H. Lubis, L. P. A. Lubis, and Sriani, "Sentiment analysis on twitter about the death penalty using the support vector machine method," *TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika*, vol. 11, no. 2, pp. 312–321, 2024, doi: 10.37373.

[8] Y. Ansori and K. F. H. Holle, "Perbandingan Metode Machine Learning dalam Analisis Sentimen Twitter," *justin*, vol. 10, no. 4, p. 429, Dec. 2022, doi: 10.26418/justin.v10i4.51784.

[9] M. Furqan, M. Ikhsan, and R. Aini, "Algoritma Support Vector Machine Untuk Analisis Sentimen Masyarakat Indonesia Terhadap Pandemi Virus Corona Di Media Sosial," vol. 4, no. 4, 2023.

[10] Kevin Perdana, Titania Pricillia, and Z. Zulfachmi, "Optimasi Textblob Menggunakan Support Vector Machine Untuk Analisis Sentimen (Studi Kasus Layanan Telkomsel)," *bangkitindonesia*, vol. 10, no. 1, pp. 13–15, Mar. 2021, doi: 10.52771/bangkitindonesia.v10i1.120.

[11] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," *JJEEE*, vol. 5, no. 1, pp. 32–35, Jan. 2023, doi: 10.37905/jjeee.v5i1.16830.

[12] A. P. Nardilasari, A. L. Hananto, S. S. Hilabi, T. Tukino, and B. Priyatna, "Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM Pada Media Sosial Twitter," *JOINTECS*, vol. 8, no. 1, p. 11, Mar. 2023, doi: 10.31328/jointecs.v8i1.4265.

[13] Friska Aditia Indriyani, Ahmad Fauzi, and Sutan Faisal, "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan support vector machine," *tekno*, vol. 10, no. 2, pp. 176–184, Jul. 2023, doi: 10.37373/tekno.v10i2.419.

[14] A. Putra and R. Latifah, "Analisis Sentimen Pengguna Twitter Terhadap Aplikasi Pinjaman Online Menggunakan Metode Support Vector Machine," *Seminar Nasional Penelitian LPPM UMJ*, 2022.

[15] A. Nofandi, N. Y. Setiawan, and D. W. Brata, "Analisis Sentimen Ulasan Pelanggan dengan Metode Support Vector Machine (SVM) untuk Peningkatan Kualitas Layanan pada Restoran Warung Wareg," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 1, pp. 458–466, 2023.

[16] I. S. Milal, M. Hasanuddin, M. A. N. Azhari, R. A. Nugraha, and N. Agustina, "KLASIFIKASI TEKS

REVIEW PADA E-COMMERCE TOKOPEDIA MENGGUNAKAN ALGORITMA SVM," *NARATIF : Jurnal Ilmiah Nasional Riset Aplikasi dan Teknik Informatika*, vol. 5, no. 1, 2023, doi: https://doi.org/10.53580/naratif.v5i1.191.

KHAIRUNNISA is computer science student at UINSU who started her undergraduate education in 2020. This research is a form of final assignment to complete his undergraduate study period.

SRIANI, is a lecturer at the Faculty of Science and Technology at UINSU. Who is also the supervisor of Azizah Oktarina Ritonga in completing this final project research. Sriani has more than 20 articles related to computer science.