CLASSIFICATION OF INSTAGRAM CONTENT BASED ON COMMENTS USING SUPPORT VECTOR MACHINE

Daniar Nur Amin[1], Ednawati Rainarli[2]

[1,2] Informatics Engineering – Indonesian Computer University
Jl. Dipatiukur 112-114 Bandung
E-mail: daniarnuramin@email.unikom.ac.id[1], ednawati.rainarli@email.unikom.ac.id[2]

ABSTRACT

Instagram is a popular social media application used by many people around the world, one of which is in Indonesia. Based on its function, this application is used to share photos and videos with other users. With the increasing number of users, posts on Instagram need to be grouped by category. This study uses comments to classify image posts on Instagram. The Support Vector Machine method was chosen to classify the text in the comments. At the preprocessing stage, there is something different from the research that has been done, namely at the stage of language normalization because at this stage the non-standard word is changed into a standard word in accordance with the KBBI. The test was carried out using the Support Vector Machine method with a linear kernel, the test was carried out with 200 training data and for the test data as many as 100 data, and the accuracy obtained from the test was 96%. Based on these results it is concluded that Support Vector Machine can be used to classify Instagram content based on comments.

Keywords: Instagram, Posts, Instagram Content, Classification, Support Vector Machine

1. INTRODUCTION

Instagram is a popular social media application used by many people around the world, one of which is in Indonesia. Based on its function, this application is used to share photos and videos with other users. With the increasing number of users, the posts in Instagram need to be grouped by category [1]. From every post on Instagram, there are many comments made by fellow users. The comments in this post have the potential to be used to group images sent into these categories. Therefore we need a way to classify Instagram content in the form of image posts based on comments.

Research on classification or grouping in the form of Indonesian language text data has been done before by Dio Ariandi, regarding the classification of Indonesian news using the Naïve Bayes Classification method and Support Vector Machine. The results were obtained using the Support Vector Machine method 88.1% better 5.9% than NBC with an accuracy of 82.2% [2]. Other researchers classify online news using the Support Vector Machine and K-Nearest Neighbor method by Siti Nur Aisyah. The accuracy of the Support Vector Machine method was 93.2% 33.2% better compared to K-Nearest Neighbor with 60% accuracy [3]. Research that has been done using the Support Vector Machine method has a good accuracy value, and therefore in this study will use the Support Vector Machine method to help classify the data that has been determined.

The Support Vector Machine method is a technique for making predictions, both in the case of classification and regression, this study will implement the Support Vector Machine method for classifying Instagram content based on comments on a post.

1.1 Classification

Classification is the process of grouping objects that have the same characteristics or characteristics into several texts [4]. Instagram content based on comments is a process for classifying or grouping data in the form of comments into certain categories, which is related to text mining. Understanding of text mining itself is the process of finding information or new trends that were not previously revealed by processing and analyzing large amounts of data. In analyzing some or all of the unstructured text, text mining tries to associate one part of the text with another based on certain rules. The expected result is new information or "insight" that was not revealed before[5][6].

1.2 Preprocessing

Preprocessing the first step taken is the training process. The training data training process will go through several main processes, namely the process of preprocessing, weighting and Support Vector Machine training. The preprocessing process itself consists of six processes that is *Case Folding, Cleansing, Filtering, Tokenizing, Normalisasi Bahasa* dan *Stopwords Removal*.

1.2.1 *Case Folding*

Case Folding is the process of homogenizing the form of words in a comment into lowercase or uppercase letters. In this study uniformed to form lowercase [6].

1.2.2 *Cleansing*

Cleansing is a process of cleaning up words that are not needed to reduce noise. The words that are omitted are URL, hashtag (#) and username (@) [7].

1.2.3 *Filtering*

Filtering is a process of eliminating characters other than letters a to z. at this stage characters other than letters are removed [8].

1.2.4 *Tokenizing*

Tokenizing is the sentence cutting stage based on each word that makes it up. At this stage, the description that was originally in the form of sentences becomes a single word [9].

1.2.5 *Language Normalization*

Language Normalization performed on non-standard words, this stage aims to restore the writing form of each word following the Big Indonesian Dictionary. This process is done by matching each word in the training data document and test data with the words in the predefined word list[7][10].

1.2.6 *Stopwords Removal*

Stopword Removal that is the process of removing the words contained in the stopword list. The words in the stopword list are pronouns, connectors and pointers [3][11].

1.3 Weighting TF-IDF

Term Frequency–Inverse Document Frequency used to determine the value of the appearance of each word of documents, this calculation uses a formula to determine the level of truth of a word that is in the document [12].

Weighting is obtained from the number of words or terms that appear in each document that is commonly called the term frequency (tf). tf-idf is successfully used in filtering in various fields, including text summarization and classification [12].

The weight of a term is greater if the term often appears in a document and smaller if the term appears in many documents [12].

The idf value of a term (word) can be calculated using equation (1) following:

$$idft = \log\,(\mathrm{N}/df) \qquad (1)$$

To calculate the weight (W) of each document to the term system (words) can use the following equation (2). $Wt{=}tfdt{*}idft$

$$W\mathrm{dt} = tfdt * idft \qquad (2)$$

1.4 *Support Vector Machine*

Support Vector Machine (SVM) is a learning system that uses hypothetical spaces in the form of linear functions in a high-dimensional feature space, trained with learning algorithms that are based on the theory of optimization by implementing learning biases derived from statistical learning theory. SVM is a relatively new technique compared to other techniques but has better performance in various fields of application such as bioinformatics, handwriting recognition, text classification, prediction and so on. [13][14]. The learning process in SVM aims to get the hypothesis in the form of the best hyperplane. The best hyperplane can not only separate data but also have the largest margin. Data that is on a hyperplane is called support vector.

Data in the input space has dimension d denoted by $x_i = \in \Re^d$ while the class label is denoted by $y_i \in \{-1, +1\}$ to i = 1, 2, …nWhere n is the amount of data. It is assumed that both classes -1 and +1 can be linearly separated in the boundary plane [15], then the field 1 equation is defined in the following equation (3):

$$w.x_i + b = 0 \qquad (3)$$

Data x_i which is divided into two classes, which belong to class -1 (negative sample) is defined as a vector that satisfies the following inequality (4):

$$w.x_i + b < 0 \;\; untuk\, y_i = -1 \qquad (4)$$

While those belonging to the +1 class (positive sample) fulfill the following inequality (5):

$$w.x_i + b > \; 0\; untuk\, y_i = +1 \qquad (5)$$

Where:
x_i = input data
y_i = the label of each data
w = value of the normal field
b = the position of the plane relative to the center of the coordinates

The parameters w and b are the parameters whose values will be searched. When data labels y_i = -1, then the delimiter becomes equation (6) below:

$$w.x_i + b \;\leq\; -1 \qquad (6)$$

when labeling data y_i = +1, then the delimiter becomes equation (7) follows:

$$w.x_i + b \;\geq\; +1 \qquad (7)$$

The largest margin can be found by maximizing the distance between the boundary fields of the two classes and their closest point, which is 2/|w|. This is formulated as a quadratic programming (QP) problem that is finding the minimum point of equation (8) by paying attention to equation (9) below:

$$min\; \tau(w) = \frac{1}{2}||w||^2 \qquad (8)$$

$$y_i(w * x_i + b) - 1 \geq 0,\; (i = 1,..,n) \qquad (9)$$

This problem can be solved by sharing computational techniques. Easier to solve by changing equation (8) into the Lagrangian function in equation (10), and simplifying it into the following equation (11):

$$L(w,b,a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i(y_i((w^T x_i + b) - 1)) \quad (10)$$

$$L(w,b,a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i y_i(w^T x_i + b) + \sum_{i=1}^{n} a_i \quad (11)$$

Where a_i is a Lagrange multiplier which is zero or positive ($a_i \geq 0$). The optimal value of equation (12) can be calculated by minimizing L concerning w, b and a. It can be seen in equations (13) to (14) below:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} a_i\, y_i x_i = 0 \quad (12)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} a_i\, y_i = 0 \quad (13)$$

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{n} a_i y_i (w^T x_i + b) - \sum_{i=1}^{n} a_i = 0 \quad (14)$$

Where n is the amount of data that becomes a support vector. Because the Lagrange Multiplier (α) value is unknown, the above equation cannot be solved directly to get w and b. To solve this problem, modify Equation (10) above to be the case of maximizing with optimal conditions for duality using the KKT constraint (Karush-Kuhn-Tucker) as follows:

Syarat 1: $\quad a_i\,[y_i(w.x_i + b) - 1] = 0 \quad (15)$

Syarat 2: $\quad a_i\ \geq 0, i = 1,2,\dots,n \quad (16)$

Then the Lagrange problem for classification can be stated in equation (17) below:

$$Min\ L(w,b,a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i y_i (w^T x_i + b) - \sum_{i=1}^{n} a_i \quad (17)$$

The equation model (15) above is a Lagrange primal model. Whereas by maximizing L with respect to a_i, the equation becomes equation (18) follows:

$$L_D = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1,j=1}^{n} a_i a_j y_i y_j x_i x_j \quad (18)$$

Syarat 1: $\quad \sum_{i=1}^{n} a_i y_i = 0 \quad (19)$

Syarat 2: $\quad a_i\ \geq 0, i = 1,2,\dots,n \quad (20)$

With $x_i x_j$ is a *dot-product* erplane (decision boundary or separator) is obtained by the following equation (21):

$$(\sum_{i=1}^{n} a_i y_i x_i . z) + b = 0 \quad (21)$$

n is the amount of data, x_i it is *support vector*, z is the data to be tested and determined by class, and $x_i z$ is *inner-product* between x_i dan z. For the value of b obtained from Equation (15) in the support vector. Because αi is calculated by the numeric method and has a numerical error, the value calculated for b may not be the same. To get b. Equation (15) can be simplified into equation (22) below:

$$b_i = 1 - y_i(w.x_i) \quad (22)$$

SVM is usually used for classification problems with only 2 classes, then try to develop them to classify more than 2 classes. One approach method used to classify more than 2 classes is One Against All (OAA). OAA method for the case of k-class classification, find k hyperplane where k is a lot of classes and ρ is hyperplane. In this method ρ $^{(\ell)}$ tested with all data from the class ℓ labeled +1, and all data from other classes labeled -1 [8]. The following is an illustration for the classification problem with three classes, three binary SVMs are used in Table 1 and their use in classifying new data in Figure 1 [9].

Table 1 Sample Method One Against All

$y_i = 1$	$y_i = -1$	Hipotesis
Kelas 1	Bukan kelas 1	$f^1(x) = (w^1)x + b^1$
Kelas 2	Bukan kelas 2	$f^2(x) = (w^2)x + b^2$
Kelas 3	Bukan kelas 3	$f^3(x) = (w^3)x + b^3$

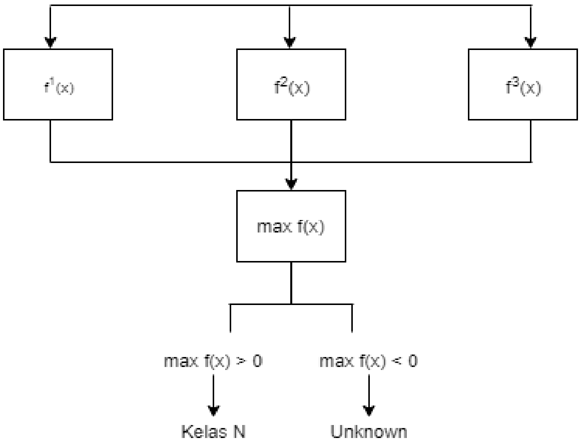


Figure 1. Method Classification SVM One Against All

The concept of the OAA approach method is assumed to be in three classes, classes 1,2 and 3. If class 1 will be tested, class 1 will be labeled +1 and class 2 and 3 will be labeled -1. If class 2 will be tested, class 2 will be labeled +1 and classes 1 and 3 will be labeled -1. And this is the same for class 3. Then you will get hyperplane for each class above. Then the class of a new data x is determined based on the largest value of the hyperplane [8]:

$$kelas\ x = arg \max_{\ell=1\dots k} ((w^{(\ell)})^T . \phi(x) + b^{(\ell)}) \quad (23)$$

1.5 Classification Performance Testing

Matrix confussion is a table that records the results of classification work. The following table is an example of a matrix confusion that classifies three-class problems [16]. Each set F_{ij} in the matrix states the number of records / data from class i whose prediction results enter class j. For example cell F_{11} is the amount of data in class 1 that is properly mapped to class 1. And F_{12} is the amount of data in class 1 that is mapped to class 2.

Table 2 Matrix Confussion

F_{ij}		Kelas Prediksi (j)		
		Kelas 1	Kelas 2	Kelas 3
Kelas Asli (i)	Kelas 1	F_{11}	F_{12}	F_{13}
	Kelas 2	F_{21}	F_{22}	F_{23}
	Kelas 3	F_{31}	F_{32}	F_{33}

The classification accuracy can be seen from the classification accuracy. Classification accuracy shows the overall performance of the classification model, where the higher the classification accuracy, the better the performance of the classification model.

$$Akurasi = \frac{jumlah\ data\ yang\ diprediksi\ secara\ benar}{jumlah\ prediksi\ yang\ dilakukan}\ x\ 100\%$$

2. CONTENTS OF RESEARCH

2.1 Problem Analysis

Based on the formulation of the problem that has been obtained that is the number of users of Instagram social networks causes many also make posts and comments. As the user will not know what posts are included in the category, if only looking at the comments that exist and must see the image or editor of the post, whereas if you want to know what the post is about just by looking at the comments there. The solution provided is the use of Machine Learning in classifying comments that are in Instagram to be categorized according to the images in the post.

2.2 Process Analysis

Process analysis is a stage for analyzing a method or analyzing the methods used. The stages used to find out what the sentence is entered into what is by the content that has been determined, in this study are divided into two stages, namely the training and testing stages. The following is an overview of the stages of the process to be carried out:

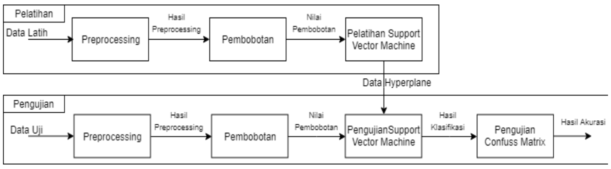


Figure 2 Process Analysis

2.3 Input Data Analysis

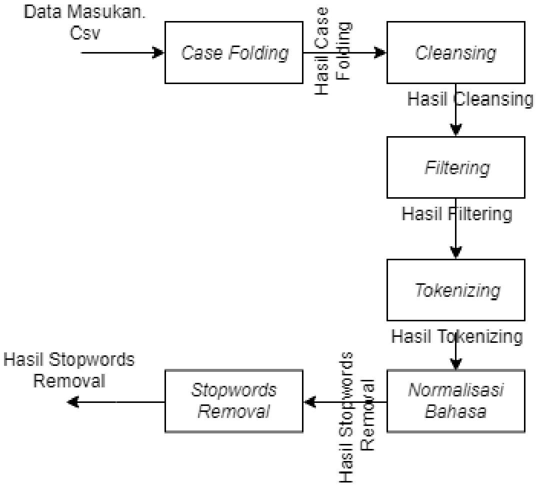
The input data used is comments on Instagram of posts related to the specified content. The comment data is obtained by using scraping techniques. The data used is from a predetermined category with CSV format.

Tabel 3 Sample Input Data

Kategori	Komentar
Game	Yap dan game nya seru ?? sangat di sayangkan versi PC nya gak ada ???? @adefirsyah namanya aja ps4 exclusive ?? @adefirsyah le Sony : eh lupa gua sabar aja gan gua juga males rilis di PC ?? Gw beli nih game wktu winter sale hrga 150K. Smpe skrang blum gw mainin krna blom di download?? Inet lemot Bangsad, gua baru beli ps4 karena gue pengen maen horizon zero dawn.. Masa sekuel terbaru nya ada di ps5?? Auto misquen gue bgsd Kasih gratis donk, biar banyak yg maen termasuk aku @zairulrizal48 bukannya kalo peli ps4 langsung dapet gratis yak? Blm bisa disebut game terbaik klo ga/blm rilis di PC ?????? @ari_arifin1 God of War (ps exclusive) game terbaik 2018

2.4 *Preprocessing* Process

Preprocessing analysis is the first stage conducted in the training process. The training data training process will go through several main processes, namely the process of preprocessing, weighting and Support Vector Machine training. The preprocessing process itself consists of six processes namely Case Folding, Cleansing, Filtering, Tokenizing, Language Normalization, and Stopwords Removal.



Gambar 4 *Preprocessing* Analysis

After carrying out the Case Folding, Cleansing, Filtering, Tokenizing, Language Normalization, and Stopwords Removal stages, the results of the preprocessing process are listed in Table 4.

Tabel 4 *Preprocessing* Results

Kategori	Hasil *Preprocessing*		
Game	game	sale	kasih
	seru	harga	gratis
	sayangkan	bermain	main
	versi	download	peli
	pc	inet	ps
	ps	lemot	gratis
	exclusive	beli	game
	le	ps	terbaik
	sony	main	rilis
	lupa	horizon	pc
	sabar	zero	god
	males	dawn	of
	rilis	sekuel	war
	pc	terbaru	ps
	game	ps	exclusive
	waktu	auto	game
	winter	miskin	terbaik

2.5 Weighting Analysis *TF-IDF*

The initial process is calculated tf (term frequency) on each document, then look for the value of df, because df is the number of documents in which a term appears.

After getting the df value, then the idf calculation is done with equation (1).

$$idf\,t = \log\,(\mathrm{N}/df) \qquad (1)$$

Taken an example in the word "game". Obtained many documents (N) = 6, and df = 2. Then, calculated as follows.

$$idf\,t = \log(6/2) = 0{,}477$$

Furthermore, to get the term weights, we calculate tf and idf with equation (2)

$$Wt = tfdt * idft \qquad (2)$$

Obtained tf = 4, and idf = 0.368. Then the calculation is as follows.

$$W\mathrm{dt} = 4 * 0{,}368 = 1{,}472$$

So the word "game" has a weight of 1.472.

2.6 Training Analysis Support Vector Machine

The matrix of each element is the result of $x_i x_j$ which will correlate with $a_i a_j$in the equation. Using the K matrix as *dot-product* $x_i x_j$in the Lagrange multiplier duality equation, is obtained:

Maximize:

$$L_D\,max = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 - \frac{1}{2}\,(53{,}486a_1^2 + 3{,}386a_1a_2 - 0{,}406a_1a_4 - 0{,}178a_1a_5 - 0{,}085a_1a_6 + 37{,}082a_2^2 - 0{,}294a_2a_3 - 0{,}043a_2a_5 - 0{,}085a_2a_6 + 64{,}723a_3^2 + 4{,}735a_3a_4 + 128{,}369a_4^2 + 0{,}677a_4a_5 + 0{,}271a_4a_6 + 48{,}275a_5^2 + 10{,}056a_5a_6 + 57{,}419a_6^2)$$

Condition 1: $a_1 + a_2 - a_3 - a_4 - a_5 - a_6 = 0$

Condition 2: $a_1, a_2, a_3, a_4, a_5, a_6 \geq 0$

In the objective function, the second term is multiplied by $y_i y_j$. he equation meets the Qudratic Programming standard so that it can be helped to solve it with a commercial solver for Quadratic Programming (QP). With the help of the software, the following results are obtained:

$$a_1 = 0.190$$
$$a_2 = 0.028$$
$$a_3 = 0.130$$
$$a_4 = 0.006$$
$$a_5 = 0.015$$
$$a_6 = 0.013$$
$$b = -61.337$$

These results indicate that all training data are support vectors due to the value of a> 0. While the value of b is obtained from the training process carried out. After finding all a and b, the SVM model can be used to predict models using:

$$f(x) = w^T.x + b$$

Where $w^T.x + b = a^T y \;\; K\big(x_i\,.x_{Uji}\big) + b.$

Then the equation $f(x)$ is as follows:

$$w^T.x + b = a^T y \;\; K\big(x_i\,.x_{Uji}\big) + b.$$

$$f(x) = \begin{bmatrix} 0.190 \\ 0.028 \\ 0.130 \\ 0.006 \\ 0.015 \\ 0.013 \end{bmatrix}^T \begin{bmatrix} 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} K(x_1.x_{Uji}) \\ K(x_2.x_{Uji}) \\ K(x_3.x_{Uji}) \\ K(x_4.x_{Uji}) \\ K(x_5.x_{Uji}) \\ K(x_6.x_{Uji}) \end{bmatrix} - 61.337$$

2.7 Analysis Testing Support Vector Machine

After finding the data value results from the calculation of $x_{Uji}x_1^T$ up to you $x_{Uji}x_6^T$. The value of the test data is substituted into equation (23) below:

$$kelas\ x = \arg\max_{k=1\ldots3}\,([w^1]^T.\varphi(x) + b^1, [w^2]^T.\varphi(x) + b^2, [w^3]^T.\varphi(x) + b^3)$$

The values of w^1 and b^1 were obtained from the results of training that had been done previously where the number 1 indicates the class 1 index, namely Games, number 2 indicates the class 2 index namely Food, number 3 indicates the class 3 index namely Sports.

$$kelas\ x = \arg\max_{k=1} \left(\begin{bmatrix} 0.190 \\ 0.028 \\ 0.130 \\ 0.006 \\ 0.015 \\ 0.013 \end{bmatrix}^T \begin{bmatrix} 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} - 61{,}337 \right)$$

$$\begin{bmatrix} 0.171 \\ 0.171 \\ -0.171 \\ -0.171 \\ -0.171 \\ -0.171 \end{bmatrix}^T \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} - 61{,}337 = 1{,}834 - 61{,}337 = -59{,}503$$

$$kelas\ x = \arg\max_{k=2} \left(, \begin{bmatrix} 0{,}007 \\ 0{,}011 \\ 0{,}022 \\ 0{,}011 \\ 0{,}008 \\ 0{,}006 \end{bmatrix}^T \begin{bmatrix} 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} + 24{,}355 \right)$$

$$\begin{bmatrix} -0{,}001 \\ -0{,}001 \\ 0{,}001 \\ 0{,}001 \\ -0{,}001 \\ -0{,}001 \end{bmatrix}^T \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} + 24{,}355 = -0{,}011 + 24{,}355 = 24{,}344$$

$$kelas\ x = \arg\max_{k=3} \left(\begin{bmatrix} 0{,}011 \\ 0{,}017 \\ 0{,}010 \\ 0{,}005 \\ 0{,}023 \\ 0{,}019 \end{bmatrix}^T \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} + 36{,}144 \right)$$

$$\begin{bmatrix} 0{,}001 \\ 0{,}001 \\ 0{,}001 \\ 0{,}001 \\ -0{,}001 \\ -0{,}001 \end{bmatrix}^T \begin{bmatrix} 5{,}688 \\ 5{,}588 \\ 0 \\ 0{,}296 \\ 0{,}085 \\ 0{,}171 \end{bmatrix} - 61{,}337 = 0{,}011 + 36{,}144 = 36{,}155$$

$$kelas\ x = \arg\max_{k=1\ldots3} (-59{,}503, \quad 24{,}344, \quad 36{,}144)$$

$$kelas\ x = 36{,}144$$

The biggest hyperplane value is 36.144 where the hyperplane value is a class 3 hyperplane value. It means that the P_{Uji} data is included in the Instagram content with the Sports category.

2.8 Testing dan Accuracy

Tests carried out to determine the accuracy of the classification of Instagram content based on comments using Support Vector Machine. Accuracy testing is a stage that has the objective to find out the accuracy value of using the Support Vector Machine method. The trick is to calculate the amount of test data that is correctly predicted by this method. And to measure performance using a confusion matrix. Testing is done into two stages, the first by testing the training data and the same test data with the amount of data as much as 200. And the second by using the training data as much as 200 and the test data as much as 100.

a. First Accuracy Testing Results

Following the first test results using training data and the same test data with the number of data 200, the accuracy values obtained can be seen in Table 5 below:

Table 5 First Accuracy Testing Results

Kondisi	Linear	RBF		
		γ=1	γ=2	γ=3
SVM	100%	100%	100%	100%
		Polynomial		
		n=1	n=2	n=3
		100%	100%	100%

The test results show that the Support Vector Machine method with all kernels shows the same accuracy of 100%.

b. Second Accuracy Testing Results

Following the results of the second test using the number of 200 training data and 100 test data, the accuracy values obtained can be seen in the following Table 6:

Table 6 Second Accuracy Testing Results

Kondisi	Linear	RBF		
		γ=1	γ=2	γ=3
SVM	96%	94%	91%	89%
		Polynomial		
		n=1	n=2	n=3
		73%	74%	74%

The test results show that the Support Vector Machine linear kernel method shows the greatest accuracy with an accuracy value of 96% compared to RBF and Polynomial. From the results of the first 96 test accuracy the test data were correctly predicted and 4 test data were incorrectly predicted.

3. CLOSING

3.1 Conclusion

Based on the results of the implementation and testing of Support Vector Machine on Instagram content based on this comment, it has been able to fulfill the purpose of the research, which is knowing the accuracy obtained by implementing the Support Vector Machine method for classifying Instagram content based on comments with the accuracy obtained for the first test of 100% and the second test is 96%. Then the conclusion can be drawn Instagram content classification based on comments by the Support Vector Machine method with Linear Kernel from the first and second tests can be done using the Support Vector Machine method. From these results, it can be seen that the Support Vector Machine method for classifying Instagram content based on comments can be used.

3.2 Suggestion

Based on the results of research in the preprocessing stages of training data and test data for comments in Instagram posts need to be considered, especially in non-standard comments, foreign languages, and regional languages. Because there are still words that are not following the Big Indonesian Dictionary (KBBI), these words need to be changed so that they are easy to process when classified.

DAFTAR PUSTAKA

[1] B. A. Kuncoro, “TF-IDF Method in Ranking Keywords of Instagram Users ’ Image Captions,” pp. 1–5, 2015.

[2] D. Ariadi and K. Fithriasari, “Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer,” *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.

[3] K. F. Siti Nur Asiyah, “Klasifikasi Berita

Online Menggunakan Metode Support Vector Machine dan K- Nearest Neighbor," *J. SAINS dan SENI ITS*, vol. 5, no. 2, pp. 317–322, 2016.

[4] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *J. Infotel*, vol. 9, no. 4, p. 416, 2017.

[5] I. Adiwijaya, "Text Mining dan Knowledge Discovery," *Kolok. bersama komunitas datamining Indones. soft-computing Indones.*, pp. 1–9, 2006.

[6] W. Athira Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," vol. 2, no. 11, pp. 4704–4713, 2018.

[7] A. A. Amrullah, A. Tantoni, N. Hamdani, R. T. R. L. Bau, M. R. Ahsan, and E. Utami, "Reviewatas Analisis Sentimen Pada Twitter Sebagai Representasi Opini Publik Terhadap Bakal Calon Pemimpin," *Pros. Semin. Nas. Multi Disiplin Ilmu Call Pap. Unisbank*, vol. 2, no. 1, pp. 978–979, 2016.

[8] L. Sofiyana, Z. Abidin, and H. Nurhayati, "Klasifikasi Emosi Untuk Teks Berbahasa Indonesia Dengan Menggunakan K-Nearest Neighbor," vol. 1, no. January, pp. 194–299, 2012.

[9] I. H. Witten, *11 - AA1*. 2002.

[10] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 5–9, 2016.

[11] F. Z. Tala, "A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia," *J. Teknosains*, vol. 6, no. 2, p. 113, 2017.

[12] D. S. Harjanto, S. N. Endah, and N. Bahtiar, "Sistem Temu Kembali Informasi pada Dokumen Teks Menggunakan Metode Term Frecency Invers Document Frequency (TF-IDF)," *J. Sains dan Mat.*, vol. 20, no. 3, pp. 64–70, 2012.

[13] E. Rainarli and A. Romadhan, "Perbandingan Simple Logistic Classifier dengan Support Vector Machine dalam Memprediksi Kemenangan Atlet," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 3, no. 2, p. 87, 2017.

[14] N. Christianini, "Support Vector and Kernel Machines." 2001.

[15] M. H, "Support Vector Machines-Kernels and The Kernel Trick," *An Elabor. Hauptseminar Read. Club Support Vector Mach.*, 2006.

[16] Eko Prasetyo, "DATA MINING - Mengolah Data menjadi Informasi Menggunakan Matlab," *Yogyakarta: ANDI*, 2014.