Non-linear Kernel Optimisation of Support Vector Machine Algorithm for Online Marketplace Sentiment Analysis

Abdul Fadlil[1*], Imam Riadi[2], Fiki Andrianto[3]

[1] *Department of Electrical Engineering, Ahmad Dahlan University, Yogyakarta, Indonesia*
[2]*Department of Information Systems, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[3]*Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*

*corr_author: fadlil@mti.uad.ac.id

Abstract: - Twitter is a social media platform that is very important in the digital world. Fast communication and interaction make Twitter a vital information center in sentiment analysis. The purpose of this research is to classify public opinion about the presence of marketplaces in Indonesia, both positive and negative sentiments, using a Non-linear SVM algorithm based on 1276 tweets. This research involves the stages of data pre-processing, labeling, feature extraction using TF-IDF, and data division into three scenarios: 80% training data and 20% test data, 50% training data and 50% test data scenario, and 20% training data and 80% test data scenario. The last process, GridSearchCV, combines cross-validation and non-linear SVM parameters for model evaluation using a confusion matrix. The best SVM model resulting from the scenario was 80% training and 20% test data, with hyperparameters Gamma = 100 and C = 0.01, achieving 89% accuracy. When tested on never-before-seen data, the accuracy increased to 90%, with an f1-score of 91%, precision of 88%, and recall of 95% on negative sentiments. In conclusion, evaluating the performance of non-linear SVM kernels with a combination of hyperparameter values can improve accuracy, especially on public response information about online marketplaces and public sentiment.

Keywords: Marketplace, SVM non-linear, Indonesia, machine learning

I. INTRODUCTION

Twitter is a powerful social media presence in the online marketplace, offering businesses a variety of opportunities to reach a broad audience and promote their products. Features such as tweets, retweets and hashtags allow sellers to build brand awareness and interact with potential customers. Twitter also serves as a valuable source of real-time information, allowing sellers to monitor market trends and gather feedback from customers. However, using Twitter effectively comes with several challenges, such as reputation management and privacy regulations. Despite these challenges, Twitter remains a powerful tool to support the growth and success of online marketplaces [1]-[2].

Twitter aside, online marketplaces have changed the business landscape by providing a digital platform for sellers to reach a larger audience and compete with larger companies. These marketplaces also benefit consumers, offering easy access to products, more excellent choices, and the ability to compare prices and read reviews. However, they also bring challenges, such as intense competition and privacy concerns. Regulations in Indonesia, such as the ITE Law and Government Regulation No. 80/2019, aim to ensure fair and orderly trading activities through online marketplaces while protecting consumer rights.

In addition, the use of Support Vector Machines (SVM) in tweet data analysis has a significant impact on understanding public opinion, trends, and reactions on social media platforms such as Twitter. SVMs enable classification and sentiment analysis of tweet data, providing valuable insights for decision-making, brand management and research purposes [3].

In previous research related to the marketplace, sentiment analysis on the shopee application using the SVM method resulted in an accuracy rate of 98% and an f1-score of 98% [4]. Analyzing shopee marketplace sentiment using the Naïve Bayes classifier, resulting in an accuracy rate of 90.03% [5]. Comparing NBC and SVM in the online marketplace, the accuracy level of SVM is 5% higher than NBC [6]. Research related to the shopee product review marketplace where the evaluation results get an accuracy rate of 90.03% using naïve Bayes [7]. Sentiment Analysis on Twitter Social Media towards Shopee E-Commerce through Support Vector Machine (SVM) Method kernel SVM accuracy rate 93.20% [8].. Penelitian berikutnya melakukan analisis fake review e-

commerce, dengan hasil akurasi terbaik pada model SVM [9]. Penelitian selanjutnya B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM dengan hasil model terbaik pada methode SVM [10].

This research will compare the performance of polynomial kernels in finding the highest accuracy value in a classification. This research is based on previous research where SVM accuracy results are better in analyzing class balance. The results of this study are to determine the best accuracy in the SVM algorithm study case of online marketplace sentiment analysis.

II. METHOD

Research processes that involve collecting and analyzing data from Twitter usually involve several vital stages that include data crawling, preprocessing, labelling, data splitting, and the use of a SVM [11]. The following flow of this research is shown in Fig. 1.

A. Crawling Data

The research flow begins with crawling data from Twitter. In this stage, tweet data is collected from Twitter using web scraping techniques or the Twitter API [12]. Data collection must be done carefully to ensure the quality and accuracy of the data to be used in the analysis.

B. Data Preprocessing

Once the tweet data has been collected, the next step is data preprocessing. This preprocessing process involves cleaning and preparing the data for further analysis. This includes removing special characters, addressing missing or duplicate data, and converting the tweet text into a format that can be used by modelling algorithms such as SVM [13]-[14]. Preprocessing can also involve text normalization and removal of irrelevant words. Preprocessing flow can be seen at Fig. 2.

There are five major stages in preprocessing to obtain data that has a good level of accuracy:

1) Case folding: the case folding process is a way of converting all text into lowercase lowercase letters, with the aim of helping to prevent errors or mismatches that may arise due to differences in letter size.

2) Text cleaning: the process of cleaning symbols, punctuation, spaces and other characters that are less relevant or noise.

3) Tokenization: changing a sentence into words or tokens.

4) Stopword removal: the process of removing or deleting words that are considered common from a text or words, often do not provide meaning or important information in a natural language analysis approach.

5) Stemming: removing affixes from words so that only the root or basic form remains.

C. Labeling

Next, the tweet data needs to be labelled. This process usually involves the attribution of a specific label or category to each tweet. For example, in sentiment analysis, tweets can be labelled as positive, negative or neutral based on their content. This labelling is important for training the SVM model in order to classify the tweets correctly [15]-[16].

In this study, the dataset derived from Twitter consists of reviews or comments that reflect a favourable point of view, opinion, or assessment of a particular subject or topic that will be labelled as positive or value 1. Meanwhile, reviews or comments that reflect an adverse point of view, opinion, or assessment or are critical of the subject or topic will be labelled as negative or value 0.

The equation for determining labeling can be seen from (1).

$$if \begin{cases} Sentiment\ Positive = 1 \\ Sentiment\ Negative = 0 \end{cases} \quad (1)$$

The flow of the labelling process can be seen in Fig. 3.

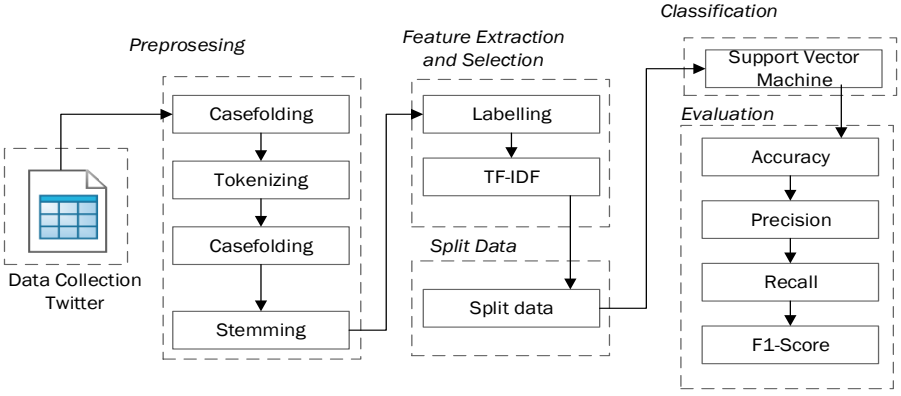

Fig. 1 Research flow

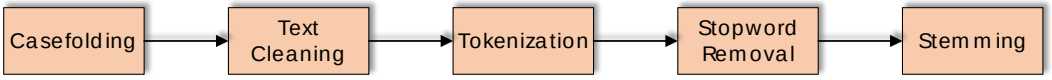

Fig. 2 Preprosesing flow

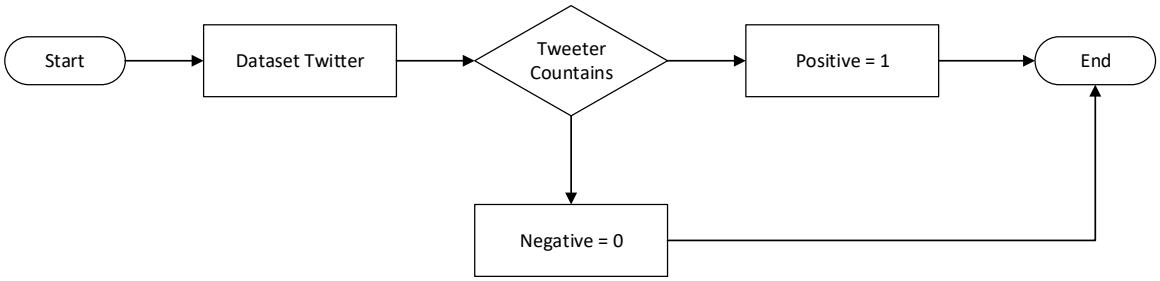

Fig. 3 Labeling flow

Tweet counters work with the help of tables that are customized to the research topic. Table I outlines the criteria used to determine the labels in the document.

As eeen in Table I, the positive label criteria include the word *gratis, beli, murah, baik, cepat, bebas, senang.* While words with negative labels include *biaya, habis, kena, capek, payah, kesel, pindah.*

D. Term weighting

Term weighting, primarily through the TF-IDF (Term Frequency-Inverse Document Frequency) method, is a technique used in text processing and information retrieval to assign appropriate weights or values to words in a document or text corpus. This technique helps in identifying the extent to which the words are relevant or necessary in a particular context [17]-[18]. Here is a more detailed explanation of TF-IDF and its formula.

Term Frequency (TF - Word Frequency): This is a metric that measures how often a word appears in a given document or text corpus.

1) Term Frequency (TF): TF is the frequency of word (t) in document (d). t describes the number of words appearing in corpus (d), and d describes the total documents in corpus (d) as in (2).

$$TF_{(t,d)} = \frac{t}{d} \quad (2)$$

TABLE I
SPECIFIC CRITERIA LABEL

Term Label Positive	Term Label Negative
Gratis	*biaya*
beli	*habis*
murah	*kena*
baik	*capek*
cepat	*payah*
bebas	*kesel*
senang	*Pindah*

2) Inverse Document Frequency (IDF): IDF is the frequency of inverted documents (t) in a corpus (d) as in (3).

$$IDF_{(t,d)} = \log \frac{t}{D} \quad (3)$$

3) TF.IDF: TF.IDF is the TF-IDF weight of word (t) in document (d) in corpus (D) as in (4).

$$TF.IDF_{(t,d,D)} = TF_{(t,d)} . IDF_{(t,D)} \quad (4)$$

Using the TF-IDF technique, words that appear frequently in documents but also appear frequently throughout the corpus will have a lower weight. In comparison, words that appear infrequently in documents but are unique in the context of the corpus will have a higher weight. This helps in finding keywords or relevant terms in text analysis, information retrieval, and various other applications in text processing and data mining [19]-[21].

E. Split Data

Data sharing involves splitting the dataset into two parts: training data and testing data. Training data is used to train the model while testing data is used to test the performance of the trained model. This aims to prevent overfitting, where the model overfits the training data and struggles to adapt to new data. This step is crucial to assess how much the model can apply knowledge from the training data to real situations [22]. In this study, there are three data split scenarios used, namely 80% training data and 20% testing data, 50% training data and 50% testing data, and 20% training data and 80% testing data. This was done to train the model and measure the level of accuracy that the model can achieve in various data-splitting contexts.

F. Support Vector Machine

SVM is a machine learning algorithm used for classification and sentiment analysis, in this case, to analyze tweet data. The model is trained using an already labelled training set and then used to classify unseen tweets in the test set. The results of using SVM can be sentiment analysis, classifying tweets into specific categories, or other tasks according to the research objectives [23]-[24]. Here are the formulas in the non-linear SVM vector space in Table II.

The accuracy of the SVM kernel model is influenced by the hyperparameters, where 'x' and 'y' indicate the two input vectors that will be projected into the higher-dimensional feature space. These hyperparameters have an important impact in determining the extent to which the model can effectively separate and accurately classify data.

G. Evaluation

In research studies, the confusion matrix helps researchers measure the accuracy and effectiveness of the classification models that researchers use. For example, in a medical study, the research may try to classify patients as positive or negative for a condition based on test results [25]-[26]. The confusion matrix will help in measuring the extent to which the model can identify patients who are actually positive (True Positive) or negative (True Negative), as well as the extent to which the model can make mistakes by classifying positive patients as unfavourable (False Negative) or vice versa (False Positive). Confusion Matrix is also used to calculate other evaluation metrics such as accuracy, precision, recall, and F1-score, all of which provide deeper insight into the performance of classification models in research studies. By using the Confusion Matrix, researchers can measure and make more accurate and reliable decisions on research results in various research fields [27]-[28]. An example of a Confusion Matrix is presented in Table III.

TABLE II
SVM KERNEL MAPPING FUNCTION

Kernel	Function
Polynomial	$K(x,y) = (x.y + c)^d$
Rbf	$K(x,y) = exp(-gamma * \|\|x - y\|\|\text{^}2)$
Sigmoid	$K(x,y) = tanh(gamma * (x * y) + c)$

TABLE III
CONFUSION MATRIX

Actual Class	***Predicted Class***	
	Positive	*Negative*
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

1) *Recall:* Calculate the success rate in identifying correct cases as correct (5).

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

2) *Accuracy:* The accuracy of the results obtained when compared with other data to assess the extent to which the correct model is accurately used (6).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

3) *Precision:* Comparison between correctly classified True Positive cases and the total predicted positive cases (7).

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

4) *F1-Score:* Balance the average weight value between precision and recall (8).

$$F1\ Score = \frac{2*recall*precision}{recall+precision} \quad (8)$$

III. RESULT AND DISCUSSION

In this report, there is a detailed description of each step that has been taken during the implementation of the research.

A. Preprosesing

Table IV describes the results of the preprocessing steps. The text in the Twitter dataset still contains characters, punctuation marks and uppercase letters, which are then converted into uniform words and lowercase letters.

Table V describes the steps in converting text into a series of word tokens. The data starts as a sentence that has been converted into lowercase letters, and then the data is broken down into tokens that represent each word in the sentence. Then, Table VI describes the steps in removing punctuation and words that lack meaning.

Table VII describes the stages in natural language processing used to remove affixes or word endings from words in the text so as to leave only the basic form or

base word [29]. Stopword removal results containing affix words are then converted into base words. The primary purpose of this process is to achieve consistency in the structure of the text and make the next stage of analysis more manageable. This process is essential in text processing efforts to analyze the information.

This preprocessing process helps ensure that the data used in analysis or modelling is of good quality, resulting in more accurate and meaningful results.

B. Labeling

Labelling uses the phyton library Textblob. This library can determine whether a text has a positive or negative sentiment. The following labelling dataset can be seen in Table VIII.

The total data of 1276 tweets consist of 538 positively labelled tweets and 738 negatively labelled tweets. Positive is symbolized by the value one, and negative is symbolized by the value 0.

TABLE IV
CASE FOLDING STEP RESULTS

Before Case Folding	After Case Folding
@ShopeeID Emang beginikah pelayanan kurir Shopee maen lempar aja dan barang tidak ada di tempat. @ShopeeID https://t.co/QBwGVsSpIc	*emang beginikah pelayanan kurir shopee maen lempar aja dan barang tidak ada di tempat*

TABLE V
TOKENIZATION STEP RESULTS

Before Tokenization	After Tokenization
emang beginikah pelayanan kurir shopee maen lempar aja dan barang tidak ada di tempat	*['emang', 'beginikah', 'pelayanan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'dan', 'barang', 'tidak', 'ada', 'di', 'tempat']*

TABLE VI
STOPWORD REMOVAL STEP RESULTS

Before Stopword Removal	After Stopword Removel
['emang', 'beginikah', 'pelayanan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'dan', 'barang', 'tidak', 'ada', 'di', 'tempat']	*['emang', 'pelayanan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'barang']*

TABEL VII
STEMMING STEP RESULTS

Before Stemming	After Stemming
['emang', 'pelayanan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'barang']	*['emang', 'layan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'barang']*

TABLE VIII
LABELING DATASET

Dataset	*Labeling*	*Conversion* Label	Total
Tweet	Positive	1	538
	Negative	0	738
Total			1276

C. Split Data

Split data is a process of dividing the dataset into two parts: training and testing data. This study divides the data into three stages, namely the split data stages of 20 training and 80 testing, 50 training and 50 testing and 80 training and 20 testing to find out which one has the best results in selecting the machine model presented in Table IX.

TABLE IX
SVM NON LINIER DATA SPLIT

		Kernel								
		Polynomial			Rbf			Sigmoid		
Data Split (%)	Training	80	50	20	80	50	20	80	50	20
	Testing	20	50	80	20	50	80	20	50	80

D. Support Vektor Machine

Svm is used to separate data into two sentiment classes, namely positive and negative, using training data on a text document that has been labelled positive and negative. Then, the model is used for new sentiment predictions on new text that the model has never seen.

Parameter grid search is a method to find the best parameter settings for an algorithm or model by testing various combinations of potential parameter values [30]. Table X shows the grid search combinations used by researchers.

Parameter grid search is used to find the best parameter settings for an algorithm or model by testing various combinations of potential parameter values. The purpose of parameter grid search is to find the combination of parameters that provide optimal model performance, measured by predefined evaluation metrics such as accuracy, precision, or recall [31]. The table presented (Table X) shows the parameter combinations used by researchers in the grid search process. For each type of kernel (Polynomial, Rbf, Sigmoid), the table shows the values tested for each parameter relevant to that kernel.

For example, for the Polynomial kernel, researchers tested combinations of values for the Degree and Coef0 parameters, with values specified within a certain range. Similarly, for the Rbf and Sigmoid kernels, researchers tested combinations of values for the Gamma, C, and Coef0 parameters, also with values specified within a certain range.

The process of determining these parameter combinations is carried out by running the model using every possible combination of parameters, then measuring the model's performance using techniques such as cross-validation or appropriate evaluation methods. By analyzing the results of each parameter combination, researchers can determine which combination of parameters produces the best model performance, thus becoming the optimal parameters to use for the model in a given case

Table XI illustrates the accuracy results, including the best parameters and highest accuracy for each non-linear SVM model. The following results are presented in the form of a bar graph as shown in Fig. 4.

TABLE X
GRID SEARCH PARAMETER

Kernel	Hyper	value
Polynomial	C	[0.1, 1, 10, 100, 1000]
	Degree	[2, 3, 4, 5, 6]
	Coef0	[0.0, 0.1, 0.5, 1.0]
Rbf	Gamma	[0.1, 0.01, 0.001, 0.0001]
	C	[0.1, 1, 10, 100, 1000]
Sigmoid	Gamma	[0.1, 0.01, 0.001, 0.0001]
	C	[0.1, 1, 10, 100, 1000]
	Coef0	[0.0, 0.1, 0.5, 1.0]

TABLE XI
GRID SEARCH PARAMETER

Kernel	Split data	Hyperparameter				
		C	Degree	Gamma	Coef0	Acurasi
Polynomial	80:20	1	2	-	0,5	0,88
	50:50	1	2	-	1,0	0,85
	20:80	1	2	-	1,0	0,78
Rbf	80:20	100	-	0.01	-	0,89
	50:50	10	-	0,1	-	0,86
	20:80	1000	-	0,001	-	0,79
Sigmoid	80:20	100	-	0,1	1,0	0,86
	50:50	10	-	0,1	0,0	0,86
	20:80	10	-	0,1	0,0	0,79

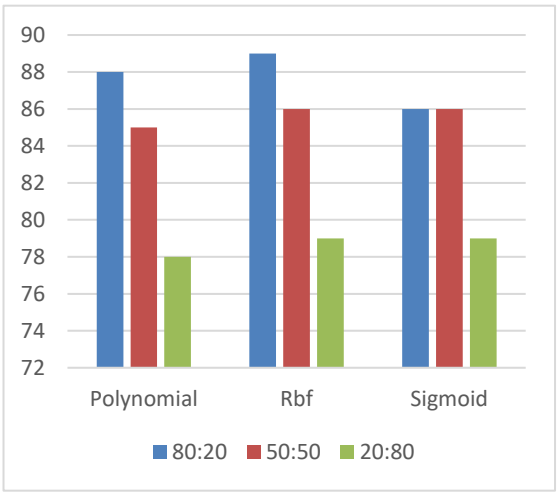

Fig. 4 A Visual analysis of an svm graphic representation

The average accuracy across all non-linear SVM kernel parameter alignments is around 89%, with the best performance at 80% training and 20% testing data split on the Rbf kernel.

E. Evaluation

Evaluation of kernel SVM models using a confusion matrix provides valuable information on the model's ability to correctly classify True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [32]. By analyzing the performance results through the confusion matrix, researchers gain an in-depth understanding of the accuracy, precision, recall, and F1-score, which provides an overall assessment of how effective the model is and also identifies possible areas for improvement. These metrics are calculated using Equations (5), (6), (7), and (8). The following results from the confusion matrix on the kernel polynomial model achieved 90% accuracy, 88% precision, 96% recall and 92% f1-score for negative sentiment detection. These results are summarized in Table XII.

The classification results of the model can be seen in Fig. 5, where it was noted that the model successfully identified TP 140, TN 91, FP 19 and FN 6 out of the total tweet data analyzed.

TABLE XII
POLYNOMIAL KERNEL TEST RESULTS

	Precision	Recall	F1-Score	Support
0	0,88	0,96	0,92	146
1	0,94	0,83	0,88	110
Accuracy			0,90	256
Macro Avg	0,91	0,89	0,90	256
Weighted Avg	0,91	0,90	0,90	256

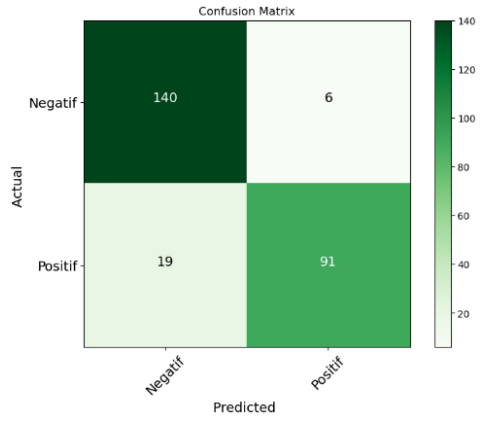

Fig. 5 Confusion matrix for the polynomial kernel

The following results from the confusion matrix on the Rbf kernel model achieve 90% accuracy, 88% precision, 95% recall and 91% f1-score for Negative sentiment detection. These results are summarized in Table XIII.

The classification results of the model can be seen in Fig. 6, where it was noted that the model successfully identified TP 139, TN 91, FP 19 and FN 7 out of the total tweet data analyzed.

The following results from the confusion matrix on the sigmoid kernel model achieve 89% accuracy, 88% precision, 95% recall and 91% f1-score for Negative sentiment detection. These results are summarized in Table XIV.

TABLE XIII
RBF KERNEL TEST RESULTS

	Precision	Recall	F1-Score	Support
0	0,88	0,95	0,91	146
1	0,93	0,83	0,87	110
Accuracy			0,90	256
Macro Avg	0,90	0,89	0,89	256
Weighted Avg	0,90	0,90	0,90	256

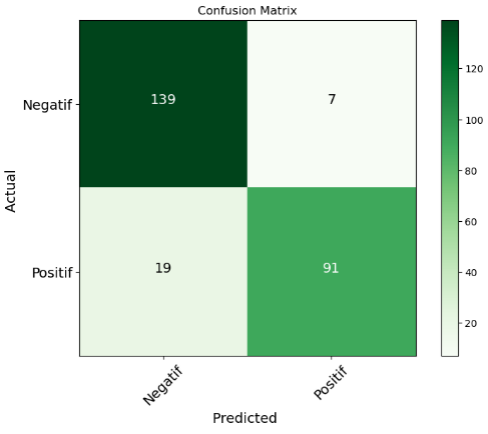

Fig. 6 Confusion matrix for the Rbf kernel

TABLE XIV
SIGMOID KERNEL TEST RESULTS

	Precision	Recall	F1-Score	Support
0	0,88	0,95	0,91	146
1	0,92	0,83	0,87	110
Accuracy			0,89	256
Macro Avg	0,90	0,89	0,89	256
Weighted Avg	0,90	0,89	0,89	256

The pre-processing steps The classification results of the model can be seen in Fig. 7, where it is noted that the model successfully identified TP 91, TN 92, FP 16 and FN 17 from the total analyzed tweet data.

Rbf kernels can significantly improve accuracy in non-linear SVMs by performing careful tuning of the right hyperparameters. By adjusting parameters such as gamma and C, SVM-RBF is able to effectively adapt its model to cope with complex and non-linear relationships between features in classification tasks, thus producing more accurate results and better fit to the given data [33]. Setting these parameters wisely through tuning techniques such as CV can help improve the performance of SVM-RBF in handling non-linear classification tasks. The result is a model that is better able to understand and adapt to complex relationships in the data, which in turn improves accuracy in classifying new data in various application contexts.

IV. CONCLUSION

Sentiment analysis of online marketplaces in Indonesia was conducted on Twitter social media with 1276 tweets data of 538 positive and 538 negative sentiments using the non-linear SVM method. This process includes data preprocessing, weighting labelling using the TF-IDF method, and data separation with three scenarios, namely 80% training and 20% Testing, 50% training and 50% Testing, and 20% training and 80% Testing. GridSearchCV combines cross-validation and non-linear SVM parameters for model evaluation using a confusion matrix. The best SVM model from the scenario results obtained the best separation on 80% training and 20% testing data with the best hyperparameter on the Rbf kernel. The optimal parameters obtained from the experimental results of the value of C = 100 and gamma = 0.01 resulted in a model accuracy of 89%. When performed on a model that has never been seen, the accuracy results increase to 90% with an f1-score value of 91%, precision of 88% and recall of 95% on Negative sentiment. In conclusion, the performance evaluation of the non-linear SVM model obtained the highest accuracy results on the Rbf kernel on sentiment towards online marketplaces. The potential to improve the performance of the model can be done by setting the hyperparameters of the non-linear SVM kernel.

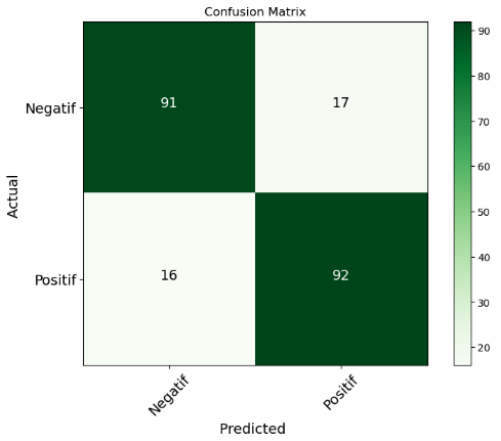

Fig. 7 Confusion matrix for the sigmoid kernel

REFERENCES

[1] L. Wang and C. A. Alexander, "Machine learning in big data," *Int. J. Math. Eng. Manag. Sci.*, vol. 1, no. 2, pp. 52–61, 2016, doi: 10.33889/ijmems.2016.1.2-006.

[2] M. I. Al-Mashhadani, K. M. Hussein, E. T. Khudir, and M. Ilyas, "Sentiment Analysis using Optimised Feature Sets in Different Facebook/Twitter Dataset Domains with Big Data," *Iraqi J. Comput. Sci. Math.*, vol. 3, no. 1, pp. 64–70, 2022, doi: 10.52866/ijcsm.2022.01.01.007.

[3] K. S. MANOJ and S. SMITA, "Support Vector Machine and Random Forest Machine Learning Algorithms for Sentiment Analysis on Tourism Reviews: a Performance Analysis," *i-manager's J. Comput. Sci.*, vol. 9, no. 3, p. 1, 2021, doi: 10.26634/jcom.9.3.18479.

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

[5] U. Makhmudah, S. Bukhori, J. A. Putra, and B. A. B. Yudha, "Sentiment Analysis of Indonesian Homosexual Tweets Using Support Vector Machine Method," *Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019*, pp. 183–186, 2019, doi: 10.1109/ICOMITEE.2019.8920940.

[6] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 10, no. 1, pp. 731–740, 2023, [Online]. Available:

https://jurnal.mdp.ac.id/index.php/jatisi/article/view/3582

[7] E. R. Kaburuan, Y. S. Sari, and I. Agustina, "Sentiment Analysis on Product Reviews from Shopee Marketplace using the Naïve Bayes Classifier," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 3, p. 150, 2022, doi: 10.24843/lkjiti.2022.v13.i03.p02.

[8] P. S. Hutapea and W. Maharani, "Sentiment Analysis on Twitter Social Media towards Shopee E-Commerce through Support Vector Machine (SVM) Method," *JINAV J. Inf. Vis.*, vol. 4, no. 1, pp. 7–17, 2023, doi: 10.35877/454ri.jinav1504.

[9] S. N. Alsubari *et al.*, "Data analytics for the identification of fake reviews using supervised learning," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[10] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *ournal Theor. Appl. Electron. Commer. Res.*, vol. 17, pp. 458–475, 2022.

[11] H. Tufail, M. U. Ashraf, K. Alsubhi, and H. M. Aljahdali, "The Effect of Fake Reviews on e-Commerce during and after Covid-19 Pandemic: SKL-Based Fake Reviews Detection," *IEEE Access*, vol. 10, pp. 25555–25564, 2022, doi: 10.1109/ACCESS.2022.3152806.

[12] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter," *J. Inf. Syst. Manag.*, vol. 3, no. 2, pp. 44–49, 2021, doi: 10.24076/joism.2021v3i2.558.

[13] A. B. Osmond and F. Hidayat, "Electronic Commerce Product Recommendation using Enhanced Conjoint Analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 666–673, 2021, doi: 10.14569/IJACSA.2021.0121176.

[14] F. El Barakaz, O. Boutkhoum, and A. El Moutaouakkil, "A new preprocessing method reduces the dimensionality of classification models," *ACM Int. Conf. Proceeding Ser.*, 2019, doi: 10.1145/3372938.3373005.

[15] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Comput. Sci.*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[16] P. Mukherjee, Y. Badr, S. Doppalapudi, S. M. Srinivasan, R. S. Sangwan, and R. Sharma, "Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection," *Procedia Comput. Sci.*, vol. 185, no. June, pp. 370–379, 2021, doi: 10.1016/j.procs.2021.05.038.

[17] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *J. Phys. Conf. Ser.*, vol. 2171, no. 1, 2022, doi: 10.1088/1742-6596/2171/1/012021.

[18] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, pp. 976–984, 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.

[19] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[20] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0192-7.

[21] L. Zhang, "Research on case reasoning method based on TF-IDF," *Int. J. Syst. Assur. Eng. Manag.*, vol. 12, no. 3, pp. 608–615, 2021, doi: 10.1007/s13198-021-01135-6.

[22] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Comput. Methods Programs Biomed. Updat.*, vol. 4, no. August, p. 100118, 2023, doi: 10.1016/j.cmpbup.2023.100118.

[23] R. Kusumawati, A. D'Arofah, and P. A. Pramana, "Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services," *J. Phys. Conf. Ser.*, vol. 1320, no. 1, 2019, doi: 10.1088/1742-6596/1320/1/012016.

[24] N. Nandal, R. Tanwar, T. Choudhury, and S. C. Satapathy, "Context driven bipolar adjustment for optimized aspect level sentiment analysis," *J. Sci. Ind. Res. (India).*, vol. 79, no. 2, pp. 122–127, 2020, doi: 10.56042/jsir.v79i2.68447.

[25] M. A. Virgananda, I. Budi, Kamrozi, and R. R. Suryono, "Purchase Intention and Sentiment Analysis on Twitter Related to Social Commerce," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 7, pp. 543–550, 2023, doi: 10.14569/IJACSA.2023.0140760.

[26] R. Yang *et al.*, "Big data analytics for financial Market volatility forecast based on support vector machine," *Int. J. Inf. Manage.*, vol. 50, no. May, pp. 452–462, 2020, doi: 10.1016/j.ijinfomgt.2019.05.027.

[27] H. Syahputra, "Sentiment Analysis of Community Opinion on Online Store in Indonesia on Twitter using Support Vector Machine Algorithm (SVM)," *J. Phys. Conf. Ser.*, vol. 1819, no. 1, 2021, doi: 10.1088/1742-6596/1819/1/012030.

[28] M. Desai and M. A. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey," *Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016*, no. March, pp. 149–154, 2017, doi: 10.1109/CCAA.2016.7813707.

[29] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, 2020, doi:

10.1109/ICDABI51230.2020.9325685.

[30] R. C. Chen and H. L. Lin, "Application of support vector machines on prediction of repeat visitation," *Int. Conf. Comput. Intell. Man-Machine Syst. Cybern. - Proc.*, vol. 1, no. April, pp. 152–157, 2006.

[31] Y. Yu *et al.*, "Quantitative analysis of multiple components based on support vector machine (SVM)," *Optik (Stuttg).*, vol. 237, no. March, p. 166759, 2021, doi: 10.1016/j.ijleo.2021.166759.

[32] A. N. Rohman, R. Luviana Musyarofah, E. Utami, and S. Raharjo, "Natural Language Processing on Marketplace Product Review Sentiment Analysis," *2020 2nd Int. Conf. Cybern. Intell. Syst. ICORIS 2020*, 2020, doi: 10.1109/ICORIS50180.2020.9320827.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning (Data Mining, Inference, and Prediction)*, vol. 26, no. 4. 1967.