KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
VOL. 21 (1) (2024), 68-82 p-ISSN: 693-7554, e-ISSN:2654-3990

Linear Kernel Optimization of Support Vector Machine Algorithm on Online Marketplace Sentiment Analysis

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

[1,3]Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 55166
[2]Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 55166
[3]Department of Information System, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 55166

Abstract

Twitter is a short message platform commonly used as a means of news information, commentary, and social interaction. One of the utilization of twitter is to analyze the sentiment of the online marketplace which can be used to determine the service, quality of goods, and delivery of goods on a product, service or application. This research aims to categorize the reviews or responses of the Indonesian people, especially to the online marketplace using the linear Support Vector Machine (SVM) algorithm. In order to make continuous improvements to the role of the Indonesian online marketplace in the future, sentiment analysis is needed. The analysis research tweets used were 4165 datasets using the python programming language. Sentiment analysis research stages include data collection, preprocessing, labeling, tf-idf weighting, split data, SVM model analysis and result evaluation. The data is then divided into 80% training data and 20% testing data, 50% training data and 50% testing data, 20% training data and 80% testing data. The results of the svm algorithm testing scenario obtained the highest optimization with an accuracy value of 97%, F1-score value on positive labels 88% and negative 98%, also obtained a positive recall value of 80% and negative 100% precision value on positive labels 98% and negative 97%, on 80% training data and 20% testing. It can be concluded that in this case, the linear svm algorithm is able to work to recognize models with a high level of accuracy so that in the future it can be used in similar cases.

Keywords: *Sentiment analysis; marketplace online; Twitter; Support Vector Machine*

1. Introduction

According to online marketplace website ranking data in July 2023, the three highest-ranked Indonesian online marketplaces are Shopee, Tokopedia, and Lazada [1]. The rapid development of online marketplaces has become a major element in consumer purchasing and interaction patterns. In Indonesia, the growth of online marketplaces has increased significantly in line with the increasingly widespread use of technology and internet access. Reviews of products and services users share on social media platforms, especially Twitter, have great potential to reflect people's views and sentiments on various aspects [2]. Twitter is included in social media, allowing users to communicate via text, video, and voice messages. The tweet is a short message, usually totaling 280 characters on Twitter social media. Twitter is often used in sentiment analysis to understand people's feelings, opinions, and attitudes about specific topics, products, brands, or events.

[*]**Corresponding author:** ***E-mail adress:*** *fiki2008048021@webmail.uad.ac.id*
Received: 25 Nov 2023, Accepted: Accepted: 23 Jan 2024 and available online 30 Jan 2024
DOI: 10.33751/komputasi.v21i1.9266

Algorithms or techniques often used in statistical data collection, machine learning, and natural language processing to analyze large and complex data are often called data mining. Data mining is a technique used by machine learning in computer learning [3]. There are machine learning algorithms that can be used, including linear SVM. Linear SVM algorithms can be applied in sentiment analysis on a review or public comment on the topic under study.

Sentiment analysis is a powerful approach to understanding public views on various topics, including online marketplaces. About the reviews or comments of the Indonesian public, especially on social media Twitter, sentiment analysis methods can provide deep insights into how consumers respond to products, services, and experiences gained from certain online marketplace platforms.

Previous researchers often use classification methods in training sentiment analysis models, including SVM. SVM includes efficient machine learning algorithms in classifying, especially in the form of text based on the results of positive and negative sentiment. Previous research conducted sentiment assessment using three labels, namely positive, negative, and neutral, in its application using the SVM Kernel RBF algorithm on Twitter application reviews [4]. Further research analyzes airline opinions using Twitter using the SVM algorithm [6]. Further research conducted sentiment analysis related to the covid 19 virus using three SVM algorithms, N-gram, and PSO, obtained the best accuracy results using the SVM algorithm [**?**]. Other related research using the SVM algorithm is to assess the sentiment of Google Meet software users utilizing the SVM algorithm; the best accuracy results reached 94% [7]. Using the SVM algorithm in sentiment assessment related to the Indonesian capital relocation scheme, the best accuracy results reached 96.68% [8]. The use of sentiment analysis in the online transportation industry with the SVM algorithm based on particle swarm optimization, obtained the results of the comparison of sentiment analysis classification algorithms that the SVM algorithm is optimal if using PSO [9]. The next research reviews the jd.id online store; the best accuracy results reach 96.4% without tf-idf weighting; if using tf-idf the best accuracy can be 98% [10]. Further research discusses the assessment of the sentiment of public figures by comparing the SVM and NBC algorithms; the research results get much better NBC performance with an accuracy value of 91.48% [11]. The next researcher conducted sentiment analysis on myim3 application users, the data was obtained from the Google Play application using the SVM algorithm, and the best accuracy results reached 87% in the training and testing test scenario (70: 30) as well as the scenario (90: 10), the best accuracy results reached 87% in the RBF kernel [12]. The next researcher conducted waste sorting using the SVM algorithm applying convolutional neural networks and obtained the best accuracy of 96.16% [13]. Further research conducted a sentiment evaluation of the mobile social security application, obtained the best accuracy results reached 96%, recall 96%, and f1-score 94% [14]. Based on the analysis of previous research, this study uses SVM to optimize classification performance on online marketplace Twitter reviews.

From the high accuracy data of previous research using the SVM algorithm above, the following research does the same thing; it's just that the object studied is different, namely by discussing the analysis of the sentiment of the online marketplace in Indonesia. The research has seven main stages: data collection, pre-processing, labeling, tf-idf, split data, SVM model classification, and result evaluation. The findings of this research can serve as a guideline to assess the optimal parameters in the use of linear SVM kernel with the division of 80% training data and 20% testing, 50% training data and 50% testing, 20% training data, and 80% testing. In addition, this research also aims to identify whether negative or positive sentiments are associated with an online marketplace platform.

2. Research Methodology

2.1. Marketplace Online

The online marketplace is a concept that is transforming modern commerce. In this digital environment, sellers from different places can offer their products and services to a broader audience, even across the globe, without having to have a physical store. On the buyer's side, online marketplaces provide the convenience of browsing a wide selection of products, reading reviews, and comparing prices from the comfort of their homes. In addition, these marketplaces often pro-

vide secure payment systems, efficient shipping, and customer support that make transactions easy [15]. This has changed how we shop and sell, making online marketplaces a critical growth engine for the digital economy. Online marketplaces also provide very lucrative business opportunities for small and medium-sized enterprises [16] and individuals who want to start a business. They can easily open their virtual stores on existing marketplace platforms, saving on high infrastructure and marketing costs. As such, online marketplaces facilitate economic growth by enabling more people to engage in e-commerce, creating healthy competition, and stimulating innovation in various sectors. With the continuous development of technology and changing consumer needs, online marketplaces are expected to remain an essential part of the global commerce ecosystem in the future. Three major Indonesian online marketplaces include Tokopedia, Shopee, and Lazada [17].

2.2. Method of collecting data

Collecting data is a decisive stage in the research, where researchers conduct sentiment analysis of community responses or tweets regarding the online marketplace on Twitter application users [11]. Data is retrieved using the phyton programming language with a librarian from the tweet harvest librarian developer. Tweets that can be retrieved from this librarian are 4165 datasets collected by researchers in August 2023 data, including date, username, and tweet.

2.3. Research stages

Research stages are a series of steps or processes in a scientific study to achieve research objectives. This stage is a stage starting from the process of retrieving datasets from Twitter through python using twitter-harvest or crawling datasets. The dataset is then subjected to removing unimportant words or preprocessing. The clean data is then subjected to Tf-idf weighting and labeling. Labeling data is then tested on predetermined training and testing scenarios [18]. The last research process is testing the classification of the SVM algorithm and evaluating the results. The stages of the SVM algorithm process in sentiment analysis research are presented through several steps described in Figure 1.

Figure 1. SVM Algorithm Process Stages

A detailed description of the research sequence or procedure is shown in Figure 1:

1. Twitter Crawling

A detailed description of the research sequence or procedure is shown in Figure 1: Data crawling is the process of retrieving data from certain sources or media for the purpose of performing analysis, including sentiment analysis [19][20]. In the context of sentiment analysis, data crawling can be used to collect large amounts of text data, such as tweets on Twitter. The researcher's goal is to understand the feelings and opinions contained in tweet data. This research uses the Python programming language to retrieve tweet data. The successfully retrieved research data amounted to 4165 data from tweets with keywords related to the Online Marketplace, such as "Tokopedia reviews," "Lazada reviews," and "Shopee reviews."

2. Preprocessing

Preprocessing is the first step in a data analysis that is done to present rough or raw data so that it can be processed further. Preprocessing aims to improve data quality and remove disturbances or anomalies that might affect the analysis results. This process includes tasks such as the removal of missing or irrelevant data, data normalization, and noise removal to produce a more consistent and relevant dataset. Data preprocessing is an important step in preparing data for further analysis. In preprocessing, the stages of text preprocessing processes, such as case folding, tokenization, stopwords removal, normalization, and stemming are described with illustrations in Figure 2. The explanation of the stages of the text preprocessing process starts from the case folding stage or changing the uppercase letters in a sentence so that the sentence form is uniform. The next flow is tokenization, which is changing the sentence into a smaller word form [21]. The next flow is

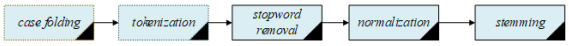

Figure 2. Stages of text preprocessing

stopword removal, which aims to eliminate common words in the text that often appear but do not provide important or relevant information in text analysis. The next flow normalization aims to simplify the text and convert the text into a uniform or standard form so that it is easier to process and analyze. Stemming is a step in natural language processing that aims to convert words into a standardized form by removing word endings [22].

3. Labeling

TextBlob is a Python labeling library used for Natural Language Processing (NLP). One of the features provided by TextBlob is the ability to perform text labeling in narratives. Text labeling in TextBlob generally refers to the act of assigning categories or additional information to text based on the context or meaning of the text. It can be used in various NLP applications, such as sentiment analysis, context-based entity recognition, Named Entity Recognition (NER), and text clustering [23].

4. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a method of assigning weight values to words in a document or text corpus used by researchers to evaluate the extent to which the word or term is important . By combining the frequency of occurrence of a word (TF) in a particular document and the uniqueness of that word across the corpus (IDF), TF-IDF assigns higher weighting values to words that frequently appear in research documents, TF-IDF is used for various purposes, such as relevant keyword discovery, document classification, or sentiment analysis. TF-IDF helps researchers to identify and understand the meaning and relevance of words in the text, thus allowing them to gain deeper insights and make more informed decisions in the context of their research [24]. The formula for calculating the weight value of words in a document is shown in calculations 1 and 2.

$$TF - IDF_{(t,d)} = TF_{(t,d)}\, x\, IDF_{(t)} \tag{1}$$

$$IDF_{(t)} = log(N/df_{(t)}) \tag{2}$$

Where:

$TF_{(t,d)}$ is the number of specialized words or terms that appear in a document.

$IDF_{(t)}$ is the ratio of the total number of documents in the corpus to the number of documents containing that term.

IDF(t) is a calculation using $\log(N/df_{(t)})$ where N is the total value of documents in the data set and $df_{(t)}$ is the value of the number of documents containing the word.

5. Training and testing model

Data sharing in modeling is used to test model performance with data that has never been seen before so that researchers can measure the extent to which the model can generalize from training data to data that has never been seen [25]. Training data is a portion of the dataset that is used to train machine learning models. The model utilizes the data to understand patterns, relationships, and distinctive properties in the dataset in order to predict or make decisions. The training data must cover a wide range of variations in the dataset as a whole for the model to gain a strong understanding. Training data is used to measure the extent to which the model is able to extract common patterns or important features from the training data and apply them well to new data. This data reflects the ability of a model to provide accurate predictions or results on various situations or data that have never been seen before.

6. Algoritma Support Vector Machine

Linear SVM is a machine learning algorithm used for regression and classification tasks. SVM allows the separation of data classes by finding the hyperplane with the maximum margin, which is the largest distance between the hyperplane and the shortest data points of each class. In addition, SVM can also be used for regression, predicting continuous values [26]-[28]. The choice of kernels, such as polynomial kernel, linear kernel, and Radial Basis Function (RBF) kernel, allows SVM to handle non-linear data. SVM is a powerful tool in a variety of applications, including text classification, image recognition, and more.

7. Parameter Kernel Linier C

The C parameter in linear SVM is also known as the error penalty parameter. It is a value that controls how much penalty is imposed on the SVM model when misclassification occurs in the training data. In some contexts, it is also referred to as the regulation parameter because it governs the extent to which the SVM model follows the regulation rules in adjusting the decision boundary. In other words, the value of C controls the extent to which the SVM model adapts to the training data by allowing a certain amount of error or avoiding misclassification completely [29].

8. Confusion Matrix

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 [30],[31]. 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 negative (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 [32], [33]. An example of a Confusion Matrix is presented in Table 1.

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

Table 1. Confusion Matrix

Four results were obtained based on the information listed in Table 1.

a. Recall

Calculate the success rate in identifying correct cases as correct

$$Recall = TP/(TP + FN)(4)$$

b. Accuracy

The accuracy of the results obtained when compared with other data to assess the extent to which the correct model is accurately used.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)(5)$$

c. Precision

Comparison between correctly classified True Positive cases and the total predicted positive cases.

$$Precision = TP/(TP + FP)(6)$$

d. F1-Score

Balance the average weight value between precision and recall.

$$F1Score = (2 * recall * precision)/(recall + precision)(7)$$

3. Result and Discussion

3.1. Crawling Dataset

The initial stage in sentiment analysis is collecting data, which involves extracting information from various sources using specialized tools such as "crawlers" in Python coding. This data is obtained from reviews or comments from the Indonesian people, especially using social media such as Twitter related to online marketplace analysis . The results of the data collection process obtained 4165 tweets, which were then used for training and testing. The dataset of 4165 online marketplace sentiment tweets is presented in Table 2.

No	Text
1	anjritlah lazada skrg ada biaya layanan
2	hah lazada skrg ada biaya layanan:')
3	@purpletaemerr @discountfess Beli voucher kebut 3000 bayar rp10 di fs lazada tiap jam 00.00, biasanya 30 detik aja langsung habis. Nah checkoutnya di murah pol yang 1000an maks 3 kalo mau jadi gratis cuma bayar biaya layanan 1k
4	Bayar tagihan wifi pake tokped sekarang kena jasa layanan seribu hue hue udah paling bener shopee kalau ga lazada
4162	Lazada kenapa ikutan ada biaya layanan juga
4163	@ShopeeID Emang beginikah pelayanan kurir Shopee maen lempar aja dan barang tidak ada di tempat. @ShopeeID https://t.co/QBwGVsSpIc
4164	@romd$_n$ @tokopedia Asli mahal pisan Makin gede nominal barang malah makin gede biaya layanan sama aplikasi, tujuan beli di tokped pas momen Electronic sale meh murah, malah kepentok biaya layanan dan aplikasi jadi sarua keneh wkwkwkwkwk
4165	T okopedia udah ada biaya layanan mayan juga yaa

Table 2. Online MarketPlace Dataset

3.2. Data Processing

The sentiment scoring process can be used appropriately and correctly after several preprocessing processes. These steps include converting the text to lowercase, breaking the text into smaller parts referred to as tokens, removing irrelevant common words, normalizing punctuation and symbols such as emojis, and converting words into base form [29]. All these steps help in producing data that is more structured and ready for analysis. The final results of this text preprocessing stage are shown in Table 3.

No	Text
1	kualiti karpet yang terbaik dengan harga mampu milik layanan mesra open everyday from am to pm ingin dapatkan barangan keperluan rumah anda kami available di shopee tiktok shop lazada dan kedai kami di klang selangor ya
2	paket ready to ship skuy beli gelasan layangan tasik di toko gelasan tiga saudara dan gelasan opat sawargi pengiriman cepat gratis ongkir bebas biaya layanan apk
3	paket gelasan layangan ready to siap rts siap kirim sudah lk paket dominasi pengiriman cepat gratos ongkir bebas biaya layanan
4	gua juga abis beli baju allhamdullilah banget belum ada biaya layanan amp gratis ongkir juga loh emang the best banget lazada
4162	lazada kenapa ikutan ada biaya layanan juga
4163	emang beginikah pelayanan kurir shopee maen lempar aja dan barang tidak ada di tempat
4164	asli mahal pisan makin gede nominal barang malah makin gede biaya layanan sama aplikasi tujuan beli di tokped pas momen electronic sale meh murah malah kepentok biaya layanan dan aplikasi jadi sarua keneh wkwkwkwkwk
4165	tokopedia udah ada biaya layanan mayan juga yaa

Table 3. Data Resulst Prerocessing text

3.3. Labeling

In this step, the tweets processed through the long preprocessing step will be labeled according to the comments using the TextBlob library in the Python programming language. This labeling library is used to build models and predict new cases. Details of the total dataset information used are presented in Table 3.

Dataset	Labeling	Conversion Label	Total
Tweet	Positive	1	538
	Negative	0	3627
Total			4165

Table 4. Dataset Labeling

The total dataset obtained is 4165 tweets with positive labeling data of 538 with the determination of conversion label one and negative data of 3627 tweets with the determination of conversion label 0. Labeling online marketplace data can be explained in Table 4.

No	Text	Text Stemmed	Labeling
1	kualiti karpet yang terbaik dengan harga mampu milik layanan mesra open everyday from am to pm ingin dapatkan barangan keperluan rumah anda kami available di shopee tiktok shop lazada dan kedai kami di klang selangor ya	['alit', 'karpet', 'baik', 'harga', 'milik', 'layan', 'mesra', 'open', 'everyday', 'from', 'am', 'to', 'pm', 'dapat', 'barangan', 'perlu', 'rumah', 'available', 'shopee', 'tiktok', 'shop', 'lazada', 'kedai', 'klang', 'selangor', 'ya']	Positif
2	paket ready to ship skuy beli gelasan layangan tasik di toko gelasan tiga saudara dan gelasan opat sawargi pengiriman cepat gratis ongkir bebas biaya layanan apk	['paket', 'ready', 'to', 'ship', 'skuy', 'beli', 'gelas', 'layang', 'tasik', 'toko', 'gelas', 'saudara', 'gelas', 'opat', 'sawargi', 'kirim', 'cepat', 'gratis', 'ongkir', 'bebas', 'biaya', 'layan', 'apk']	Positif
3	paket gelasan layangan ready to siap rts siap kirim sudah lk paket dominasi pengiriman cepat gratos ongkir bebas biaya layanan	['paket', 'gelas', 'layang', 'ready', 'to', 'rts', 'kirim', 'lk', 'paket', 'dominasi', 'kirim', 'cepat', 'gratos', 'ongkir', 'bebas', 'biaya', 'layan']	Positif
4	gua juga abis beli baju allhamdullilah banget belum ada biaya layanan amp gratis ongkir juga loh emang the best banget lazada	['gua', 'abis', 'beli', 'baju', 'allhamdullilah', 'banget', 'biaya', 'layan', 'amp', 'gratis', 'ongkir', 'loh', 'emang', 'the', 'best', 'banget', 'lazada']	Positif
4162	lazada kenapa ikutan ada biaya layanan juga	['lazada', 'ikut', 'biaya', 'layan']	Negatif
4163	emang beginikah pelayanan kurir shopee maen lempar aja dan barang tidak ada di tempat	['emang', 'layan', 'kurir', 'shopee', 'maen', 'lempar', 'aja', 'barang']	Negatif
4164	asli mahal pisan makin gede nominal barang malah makin gede biaya layanan sama aplikasi tujuan beli di tokped pas momen electronic sale meh murah malah kepentok biaya layanan dan aplikasi jadi sarua keneh wkwkwkwkwk	['asli', 'mahal', 'pis', 'gede', 'nominal', 'barang', 'gede', 'biaya', 'layan', 'aplikasi', 'tuju', 'beli', 'tokped', 'pas', 'momen', 'electronic', 'sale', 'meh', 'murah', 'kepentok', 'biaya', 'layan', 'aplikasi', 'sarua', 'keneh', 'wkwkwkwkwk']	Negatif
4165	tokopedia udah ada biaya layanan mayan juga yaa	['tokopedia', 'udah', 'biaya', 'layan', 'mayan', 'yaa']	Negatif

Table 5. Your Table Caption

3.4. Support Vector Machine and Confusion Matrix

Data labeled positive and negative is then weighted in numerical form so that the classification model can read it for each word. Furthermore, the data is processed to weigh the numerical value of each word using the tf-idf method in Python using the TfidfTransformer library [25]. Furthermore, divide the data into 80

The accuracy results on the test value of the research parameter testing 20% parameter c with a value of 0.1 obtained a value of 0.89, value 0.01 obtained a value of 0.87, value 0.05 obtained a value of 0.87, value 0.25 obtained a value of 0.90, value 0.5 obtained a value of 0.92, value 0.75 obtained a value of 0.94, value 1 obtained a value of 0.94, and the highest value in test 10 obtained a value of 0.96.

The accuracy results on the test value of the research parameter testing 50% parameter c with

No	Value	Data Testing		
		20%	50%	80%
1	0,1	0,89	0,88	0,88
2	0,01	0,87	0,88	0,88
3	0,05	0,87	0,88	0,88
4	0,25	0,90	0,90	0,89
5	0,5	0,92	0,91	0,90
6	0,75	0,94	0,92	0,91
7	1	0,94	0,92	0,91
8	10	0,96	0,94	0,92

Table 6. Parameter c Linear Kernel

a value of 0.1 obtained a value of 0.88, value 0.01 obtained a value of 0.88, value 0.05 obtained a value of 0.88, value 0.25 obtained a value of 0.90, value 0.5 obtained a value of 0.92, value 0.75 obtained a value of 0.92, value 1 obtained a value of 0.92, and the highest value in test 10 obtained a value of 0.94.

The accuracy results on the test value of 80% testing research parameter c with a value of 0.1 obtained a value of 0.88, value 0.01 obtained a value of 0.88, value 0.05 obtained a value of 0.88, value 0.25 obtained a value of 0.89, value 0.5 obtained a value of 0.90, value 0.75 obtained a value of 0.91, value 1 obtained a value of 0.91, and the highest value in test 10 obtained a value of 0.92.

The results obtained from the research that includes adjusting the parameter c in the linear SVM kernel are presented as a split data testing graph in Figure 3.

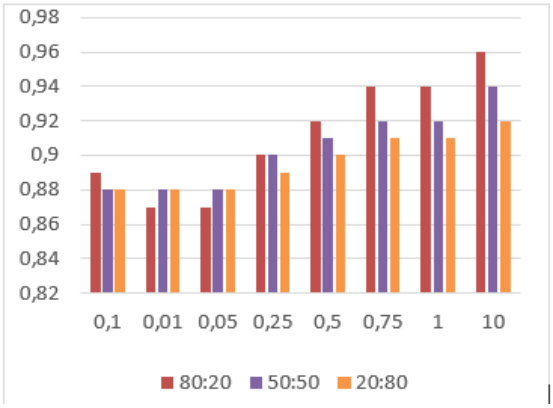

Figure 3. Split Data Testing Chart

Confusion matrix results of kernel parameter c testing calculation of the best SVM algorithm with 80% training data and 20% testing are presented in Figure 4. SVM modeling with 80% training and 20% testing data is presented in Table 6.

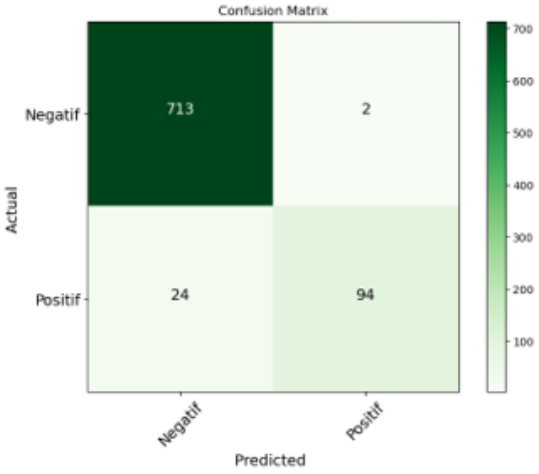

Figure 4. Confusion matrix with 80% Training and 20% Testing Data Value

	Precision	Recall	F1-Score	Support
0	0,97	1,00	0,98	715
1	0,98	0,80	0,88	118
Accuracy			0,97	833
Marco AVG	0,97	0,90	0,93	833
Weighted AVG	0,97	0,97	0,97	833

Table 7. Confusion Matrix

The best linear SVM algorithm model testing accuracy is 97%, with an F1-score value on positive labels of 88% and negative 98%; it also obtained a positive recall value of 80% and a negative 100% Precision value on positive labels of 98% and negative 97%, with a total of 833 test data. It can be concluded that the research results get the value of TP = 713, TN = 94, FP = 24, and FN = 2.

$$Accuracy = (713 + 94)/(713 + 24 + 2 + 94)$$

$$Accuracy = 807/833$$

$$Accuracy = 0.97$$

Confusion matrix results of kernel parameter c testing calculation of the best SVM algorithm with 50% Training and 50% Testing data are presented in Figure 5. Confusion matrix results of kernel parameter c testing calculation of the best SVM algorithm with 50% Training and 50% Testing data are presented in Table 7.

	Precision	Recall	F1-Score	Support
0	0,96	0,99	0,98	1798
1	0,94	0,76	0,84	285
Accuracy			0,96	2083
Marco AVG	0,95	0,88	0,91	2083
Weighted AVG	0,96	0,96	0,96	2083

Table 8. Confusion Matrix

The best linear SVM algorithm model testing accuracy is 96% with an F1-score value on positive labels of 84% and negative 98%, also obtained positive recall value of 76% and negative

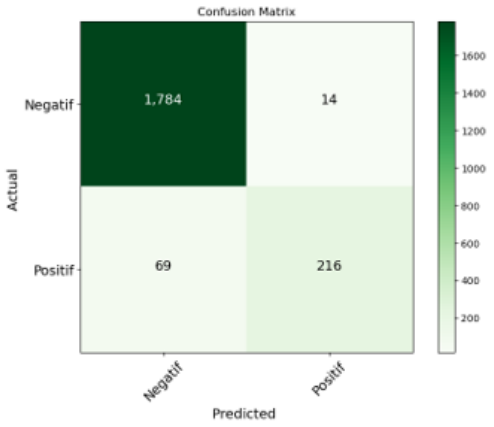

Figure 5. Confusion matrix with 50% Training and 50% Testing Data Value

99% Precision value on positive labels of 94% and negative 96%, with a total of 2083 test data. It can be concluded that the research results get the value of TP = 1784, TN = 216, FP = 69, and FN = 14

$$Accuracy = (1784 + 216)/(1784 + 69 + 14 + 216)$$

$$Accuracy = 2000/2083$$

$$Accuracy = 0.96$$

Confusion matrix results of kernel parameter c testing calculations from the best SVM algorithm with the amount of testing data of 80% are presented in Figure 6. Confusion matrix results of

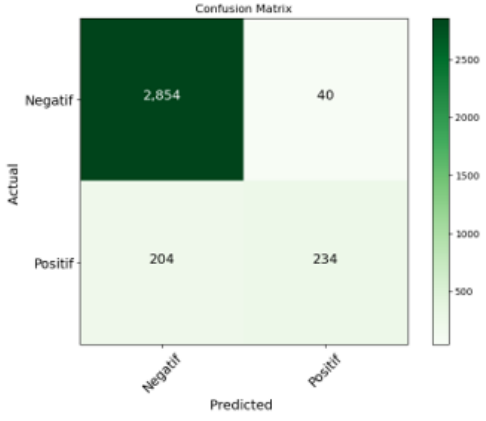

Figure 6. . Confusion matrix with 20% Training and 80% Testing Data Value

kernel parameter c testing calculation of the best SVM algorithm with 20% training data and 80% testing are presented in Table 8. The accuracy of testing the best linear SVM algorithm model is 93% with an F1-score value on positive labels of 66% and negative 96%, also obtained positive recall value of 53% and negative 99% Precision value on positive labels 85% and negative 93%, with a total of 3332 test data. It can be concluded that the research results get the value of TP = 2854, TN = 234, FP = 204, and FN = 40.

$$Accuracy = (2854 + 234)/(2854 + 40 + 204 + 234)$$

	Precision	Recall	F1-Score	Support
0	0,93	0,99	0,96	2894
1	0,85	0,53	0,66	438
Accuracy			0,93	3332
Marco AVG	0,89	0,76	0,81	3332
Weighted AVG	0,92	0,93	0,92	3332

Table 9. Classification results of models with SVM algorithm

$$Accuracy = 3088/3332$$

$$Accuracy = 0.93$$

Data(%)									
No	Training	Testing	Accuracy	F1-Score		Recall		Presision	
				Positif	Negatif	Positif	Negatif	Positif	Negatif
1	80	20	97	88	98	80	100	98	97
2	50	50	96	84	98	76	99	94	96
3	20	80	93	66	96	53	99	85	93

Table 10. Classification results of models with SVM algorithm

The results of the SVM algorithm testing scenario above, with a comparison of 80% training and 20% testing data, obtained the best model accuracy of 97%. The F1-score value on positive labels is 88% and negative 98%, also obtained a positive recall value of 80% and negative 100% The precision value on positive labels is 98% and negative 97%. At the same time, 50% training and 50% testing data obtained the best model accuracy results of 96%. The F1-score value on positive labels is 84% and negative 98%, also obtained a positive recall value of 76% and negative 99% The precision value on positive labels is 94% and negative 96%. With 20% training data and 80% testing, the best model accuracy result is 93. The F1-score value on positive labels is 66% and negative 96%, also obtained a positive recall value of 53% and negative 99% The precision value on positive labels is 85% and negative 93%.

4. Conclusion

This dataset was taken in August 2023, with 4165 datasets on Twitter social media, with keywords related to online marketplaces. The processing results of taking datasets via Twitter are then carried out at the preprocessing, labeling, and Tf-IDF weighting stages. The labeling process is done automatically using the textblob library with the Python programming language. The data is then tested on a linear SVM model with 80% training data and 20% training data, 50% training data and 50% training data, 20% training data and 80% training data. Then, the model with a confusion matrix obtained the results of the highest accuracy test scenario on 80% training data and 20% testing and obtained the best model accuracy of 97% with 833 testing data. The F1-score value on positive labels was 88% and negative 98%, also obtained a positive recall value of 80% and negative 100% Precision value on positive labels was 98% and negative 97%. Based on these results, it is concluded that the optimal data division in the linear support vector machine algorithm is shown in 80% training data and 20% testing; the best model accuracy result is 97% with 833 testing data. The F1-score value on positive labels was 88% and negative 98%, also obtained a positive recall value of 80% and negative 100% Precision value on positive labels was 98% and negative 97%.

5. References

References

[1] Similar-web, "Top Websites Ranking Most Visited Marketplace Websites in Indonesia," 2023. https://www.similar-web.com/top-websites/Indonesia/e-commerce-and-shopping/marketplace/ (accessed Aug. 15, 2023).

[2] J. Winahyu and I. Suharjo, "Aplikasi Web Analisis Sentimen Dengan Algoritma Multinomial Naïve Bayes," KARMAPATI, vol. 10, pp. 206–214, 2021.

[3] 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.

[4] R. H. Muhammadi, T. G. Laksana, and A. B. Arifa, "Combination of Support Vector Machine and Lexicon-Based Algorithm in Twitter Sentiment Analysis," Khazanah Inform. J. Ilmu Komput. dan Inform., vol. 8, no. 1, pp. 59–71, 2022, doi: 10.23917/khif.v8i1.15213.

[5] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 3, pp. 2789–2797, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[6] F. Al Isfahani and R. Mubarok, "Analisis Sentimen Pengguna Twitter Terhadap Kebijakan Pemberlakuan Pembatasan Sosial Berskala Besar (Psbb) Dengan Metode Naïve Bayes," siliwangi, vol. 7, no. 1, pp. 19–24, 2021.

[7] D. A. Fitri and A. Putri, "Analisis sentimen pengguna aplikasi googlemeet menggunakan algoritma support vector machine," CoSciTech (Jurnal Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 472–478, 2022.

[8] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (Svm)," urnal Teknol. Inf. dan Ilmu Komput., vol. 8, no. 1, pp. 147–156, 2021, doi: 10.25126/jtiik.202183944.

[9] V. Kevin, S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization (Online Transportation Sentiment Analysis Using Support Vector Machine Based on Particle Swarm Optimization)," J. Nas. Tek. Elektro dan Teknol. Inf., vol. 9, no. 2, pp. 162–170, 2020.

[10] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," J. SIMETRIS, vol. 10, no. 2, pp. 681–686, 2019, [Online]. Available: https://jurnal.umk.ac.id/index.php/simet/article/view/3487/1883

[11] W. A. Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 11, pp. 4704–4713, 2018.

[12] P. Aditiya, U. Enri, and I. Maulana, "Analisis Sentimen Ulasan Pengguna Aplikasi Myim3 Pada Situs Google Play Menggunakan Support Vector Machine," JURIKOM (Jurnal Ris. Komputer), vol. 9, no. 4, p. 1020, 2022, doi: 10.30865/jurikom.v9i4.4673.

[13] M. Fahmi, A. Yudhana, and Sunardi, "Pemilahan Sampah Menggunakan Model Klasifikasi Support Vector Machine Gabungan dengan Convolutional Neural Network," J. Ris. Komputer), vol. 10, no. 1, pp. 76–81, 2023, doi: 10.30865/jurikom.v10i1.5468.

[14] V. Fitriyana, Lutfi Hakim, Dian Candra Rini Novitasari, and Ahmad Hanif Asyhar, "Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine," J. Buana Inform., vol. 14, no. 01, pp. 40–49, 2023, doi: 10.24002/jbi.v14i01.6909.

[15] 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

[16] A. Volume, "Analysis of the role online customer review in mediating online customer experience relationships to increase marketplace rating," J. Econ. Bus. Account., vol. 7, no. 1, 2023.

[17] I. Novitasari and F. Cuandra, "Analisis Faktor yang Mempengaruhi Minat Beli pada Marketplace Online di Kota Batam," J. Inform. Ekon. Bisnis, vol. 5, pp. 339–349, 2023, doi: 10.37034/infeb.v5i2.248.

[18] R. Maulana, A. Voutama, and T. Ridwan, "Analisis Sentimen Ulasan Aplikasi Mypertamina Pada Google Play Store Menggunakan Algoritma Nbc," J. Teknol. Terpadu Vol, vol. 7, no. 2, pp. 77–82, 2021, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/jtt/article/download/318/201

[19] T. T. Widowati and M. Sadikin, "Analisis Sentimen Twitter terhadap Tokoh Publik dengan Algoritma Naive Bayes dan Support Vector Machine," Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 11, no. 2, pp. 626–636, 2021, doi: 10.24176/simet.v11i2.4568.

[20] S. H. Hardi and K. D. Hartomo, "Sentiment Analysis of Simobi Plus Mobile Application Using Naïve Bayes Classification," J. media Inform. budidarma, vol. 7, no. 3, pp. 1117–1124, 2023, doi: 10.30865/mib.v7i3.6300.

[21] Murni, I. Riadi, and A. Fadlil, "Analisis Sentimen Hate Speech pada Pengguna Layanan Twitter dengan Metode Naïve Bayes Classifier (NBC)," JURIKOM (Jurnal Ris. Komputer), vol. 10, no. 2, 2023, doi: 10.30865/jurikom.v10i2.5984.

[22] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," Edutic - Sci. J. Informatics Educ., vol. 7, no. 1, pp. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.

[23] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet Mental Health," J. Pendidik. Teknol. Inf., vol. 3, no. 2, pp. 161–171, 2023.

[24] A. Z. Praghakusma and N. Charibaldi, "Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus: Komisi Pemberantasan Korupsi)," JSTIE (Jurnal Sarj. Tek. Inform., vol. 9, no. 2, p. 88, 2021, doi: 10.12928/jstie.v9i2.20181.

[25] A. S. Nugroho, R. Umar, and A. Fadlil, "Klasifikasi Botol Plastik Menggunakan Multiclass Support Vector Machine," J. Khatulistiwa Inform., vol. 9, no. 2, pp. 79–85, 2021, doi: 10.31294/jki.v9i2.11058.

[26] 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.

[27] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study," Proc. - 2020 Int. Conf. Adv. Comput. Commun. Mater. ICACCM 2020, pp. 194–199, 2020, doi: 10.1109/ICACCM50413.2020.9213011.

[28] 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.

[29] A. A. Firdaus, A. Yudhana, and I. Riadi, "Sentiment Analysis on 2024 Presidential Election Projection using Support Vector Machine Method," Decod. J. Pendidik. Teknol. Inf., vol. 3, no. 2, pp. 236–245, 2023, [Online]. Available: http://journal.umkendari.ac.id/index.php/decode

[30] Vynska Amalia Permadi, "Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura," J. Buana Inform., vol. 11, pp. 141–151, 2020.

[31] N. S. Wardani, A. Prahutama, and P. Kartikasari, "Analisis Sentimen Pemindahan Ibu Kota Negara Dengan Klasifikasi Naïve Bayes Untuk Model Bernoulli Dan Multinomial," J. Gaussian, vol. 9, no. 3, pp. 237–246, 2020, doi: 10.14710/j.gauss.v9i3.27963.

[32] R. Umar, I. Riadi, and P. Purwono, "Klasifikasi Kinerja Programmer pada Aktivitas Media Sosial dengan Metode Stochastic Gradient Descent," JOINTECS (Journal Inf. Technol. Comput. Sci., vol. 5, no. 2, p. 55, 2020, doi: 10.31328/jointecs.v5i2.1324.

[33] M. N. Muttaqin and I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," UNNES J. Math., vol. 10, no. 2, pp. 22–27, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm