Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction

Dody Indra Sumantiawan[a,*], Jatmiko Endro Suseno[b], Wahyul Amien Syafei[c]

[a] Master of Information Systems, Postgraduate School Diponegoro University
[b] Physics Department, Faculty of Science and Mathematics Diponegoro University
[c] Department of Electrical Engineering, Faculty of Engineering Diponegoro University

Received: 8 November 2022 ; Accepted: 30 January 2023

DOI: 10.21456/vol13iss1pp1-9

Abstract

Shopping activities in the online market, especially fashion trends, continue to increase with all the promo efforts offered. One of the considerations for buying products on the online market is to read reviews. Each consumer review shows the level of interest in the product. The number of negative reviews and the emergence of many varied reviews pose a problem in categorizing reviews. Sentiment analysis is a way of looking at the polarity of reviews to classify positive and negative reviews. The Support Vector Machine method and the combination of the Synthetic Minority Oversampling Technique (SMOTE) with Tomek Links are applied in this study. Classification using the Support Vector Machine method and the combination of the Synthetic Minority Oversampling Technique (SMOTE) with Tomek Links showed better results with an *Accuracy* of 0.92, *Precision* of 0.89, *Recall* of 0.89, and F1-*score* of 0.89 than without the combination of the Synthetic Minority oversampling Technique (SMOTE) with Tomek Links with an *Accuracy* of 0.68, *Precision* of 0.55, *Recall* of 0.99, and an F1-*score* of 0.71.

Keywords : Sentiment Analysis; Classification; *Support Vector Machine*; SMOTE; Tomek Links

1. Introduction

In the online market, one of the factors that influence the decision to buy a product is the availability of information (Li *et al.,* 2019). In addition to product descriptions, consumers will see other consumer reviews on the product as information in deciding to purchase a product (Wang *et al.,* 2022). Every sentence written by consumers has implied meaning and emotion about the brand offered. Through these reviews, consumers share experiences and provide information about products to new buyers. Based on the amount of data from scrapping reviews conducted in this research process throughout July 2022, the amount of data collected was 11,615 gross data (https://shopee.co.id/). After going through the preprocessing process, the clean data became 8628 data with a total of 3394 positive labeled data and 5234 negative labeled data (preprocessing data https://shopee.co.id/).

Problems arising based on scrapping data get responses that tend to be negative reviews by 60% positive reviews by 40%. The emergence of many varied reviews makes it difficult to categorize and analyze consumer reviews. So to find out the meaning of the trend of information in a large number of consumer reviews, a method is needed to process consumer review data quickly and automatically. Sentiment analysis is a method that can be used to find out the meaning of information trends. So by knowing the importance of consumer reviews, businesses can take steps and conduct market research regarding products and consumer characteristics to take steps for brand advancement.

Regarding previous research regarding the classification of sentiment analysis, there are several studies conducted using machine learning methods. Singla *et al.,* (2017) researched sentiment analysis on consumer reviews on the Amazon site by applying three Naïve Bayes methods, Support Vector Machine (SVM) and Decision Tree. The performance results show the Accuracy score of each algorithm, with the highest Accuracy score obtained by the Support Vector Machine algorithm of 81.77 (Singla *et al.,* 2017). Hadwan *et al.,* (2022) in his research conducted a sentiment classification analysis of measuring customer satisfaction for mobile applications using the Random Forest (RF) algorithm, Bagging, Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB) and the

*) Correspondence writer: dodyindrass@gmail.com

SMOTE technique. The comparison of each algorithm shows that the highest Accuracy score is obtained by the Support Vector Machine (SVM) algorithm, with an Accuracy score of 94.38% (Hadwan *et al.,* 2022). Arivoli and Sonali (2021) conducted a sentiment analysis study using the Polarity film review dataset from the Twitter API. The Naïve Bayes and Support Vector Machine classification algorithms are used in this study. The evaluation results using the confusion matrix of the two algorithms get an Accuracy score of the Naïve Bayes algorithm 76.67 and an Accuracy score of the Support Vector Machine algorithm 78.18 (Avolli and Sonali, 2021).

Some of the machine learning algorithms applied for sentiment classification in previous studies are Random Forest (RF), Bagging, Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), and Decision Tree. The Support Vector Machine (SVM) algorithm method was chosen as the sentiment classification method in this study because the Support Vector Machine (SVM) works by searching for hyperplanes as a way of maximizing the separation between classes (Sain and Purnami, 2015). In addition, the Support Vector Machine method was chosen because this research process used consumer review data (Lee *et al.,* 2022). The type of consumer review data is informal text data which has a lot of noise and does not respect grammar rules (Li *et al.,* 2019). Support Vector Machine is very critical in data processing and tends to produce better Accuracy performance than methods in previous research (Hadwan *et al.,* 2022).

The main challenge in machine learning classification is dealing with unbalanced data because unbalanced data can result in biased results toward the majority class and poor classification performance (Bin Alias *et al.,* 2021). Unbalanced data occurs due to an unequal balance in the number of positive and negative patterns or binary class labels in the data set (Sasada *et al.,* 2020). One of the oversampling methods is the synthetic minority oversampling (SMOTE) technique for unbalanced data classification using decision tree analysis. In applying the under-sampling method, borderline and noise problems will be detected by Tomek links. Tomek links work by removing negative classes as well as positive classes that have similar characteristics (Sain and Purnami, 2015).

This study applies document sentiment level analysis through reviews in the review column on the "https://shopee.co.id/" marketplace regarding fashion products. The sentiment analysis method is SVM with the SMOTE-Tomek Links combination to improve the data sampling workflow (Bin Alias et al., 2021). SVM has the advantage of determining the separation distance between classes using a support vector so that processing carried out on large enough data can be faster (Cortes and Vapnik, 1995). The Support Vector Machine was chosen because of its relatively high level of accuracy, around 80 to 90 percent, as well as the fairly easy implementation of the algorithm and flexibility, which can be combined with other methods (Borg and Boldt, 2020). This study aims to identify consumer reviews of products marketed through online marketplaces to obtain the best sentiment analysis classification model from the SVM method and the combination of SMOTE and Tomek Links models in sentiment review analysis.

2. Literature Review

2.1 Web Scraping

Web scraping is a technique to automatically get information from a site without copying it manually. Web scraping aims to find certain information and then compile it into different formats. The benefit of web scraping is that it makes the information that is scraped or retrieved more focused, making it easier to search for something (Chapelle and Eymeoud, 2022). The first web scraping process is to identify the target website, collect the URL page from which data will be extracted, make a request to this URL to get the HTML of the page, use a search tool or source code to find the data in HTML, then save the data in a JSON or CSV file or format. Other structures (Mustopa *et al.,* 2020).

2.2 Sentiment Analysis

Sentiment analysis is a computational study that analyzes related opinions, sentiments, and emotions expressed through text (Li *et al.,* 2019). Sentiment analysis creates a system that can then classify text in a document. Sentiment analysis examines opinions about a product or an event (Borg and Boldt, 2020). Sentiment analysis systems for text analysis combine natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to entities, topics, themes, and categories in sentences or phrases (Obiedat *et al.,* 2022).

2.3 Text Mining

Text mining converts unstructured text into a structured format to identify meaningful patterns and new insights by applying advanced analytical techniques. Preparing raw text documents or datasets is also known as text preprocessing. Text preprocessing functions to convert unstructured or haphazard text data into structured data (Baek *et al.,* 2020). The stages in text mining can be seen in Figure 1.

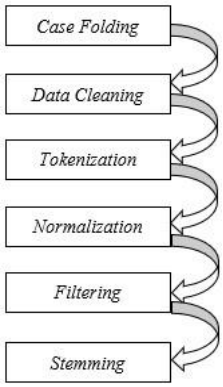

Figure 1 Stages of text mining

2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF (*Term Frequency-Inverse Document Frequency*) is a method for weighting words from the word extraction process with the idea of applying common words in *information retrieval*. This weighting method is a combination of *term frequency* and *inverse document frequency*. *Term frequency* is the number of occurrences of a term in a document. The number of terms that appear is directly proportional to the weighting given. Meanwhile, *inverse document frequency* measures how important words are in a document (Hunt, 2021).

Term Frequency (*TF*) states the number of how many *terms* that exist in a document, where the value of $f_{t,d}$ is the frequency (*f*) *term* (*t*) in document (*d*). For example, if a term for example, if a *term* is contained in a document 5 times, then the weight = 1 + *log* (5) = 1.699 is obtained. But if the term is not contained in the document, the weight is zero (0) (Hunt, 2021). The logarithmic TF formula can be seen in equation (1).

$$TF = \begin{cases} 1 + log_{10} tf_{t,d}, & if \quad tf_{t,d} > 0 \\ 0, & if \quad tf_{t,d} = 0 \end{cases} \quad (1)$$

Then next is *IDF* (*Inverse Document Frequency*), which is a calculation of how the *terms* are widely distributed in the document collection concerned. In contrast to TF, where the more frequently a word appears, the greater the value, in IDF, the less frequently a word appears in a document, the greater the value. To determine the IDF value, use the formula presented in equation (2).

$$IDF_t = log_{10} \frac{N}{df_t} \quad (2)$$

where:
df_t = The number of documents containing the *term* (*document frequency*)
N = The total number of documents

The term value will be large when the term frequently occurs (large *tf* number) but only in a few documents (large number of *idf* or small *idf*). The term value usually discards general terms. Thus the formula for the *term Weighting* (*W*) TF-IDF is a combination of the TF and IDF equations by multiplying the TF value by the IDF value, which can be seen in equation (3).

$$W_{t,d} = Wtf_{t,d} \,.\, idf_t \quad (3)$$

where:
$tf_{t,d}$ = The number of occurrences of the *term* in the document
idf = The number of documents containing the *term*

2.5 SMOTE -Tomek Links

The combination of SMOTE and Tomek Links was first introduced by Batista *et al.*, (2003), this method combines the ability of SMOTE to generate synthetic data for minority classes, and the ability of Tomek Links to remove data identified as Tomek links from the majority class, i.e., sample data from the majority class that closest to minority class data (Swana *et al.*, 2022).

2.6 Support Vector Machine (SVM)

Support vector machines are included in supervised learning, which means that the model or machine learns in advance to classify by dividing the data into two data sets, namely training data and testing data (Lee *et al.*, 2022). The *support vector machine* can find a separator function that can separate two data sets from two different classes. The concept can be explained simply that SVM tries to find the best hyperplane that functions as a separator for two classes in the input space by maximizing the distance between classes (Elhassan *et al.*, 2016). The following image of a SVM concept can be seen in Figure 2.

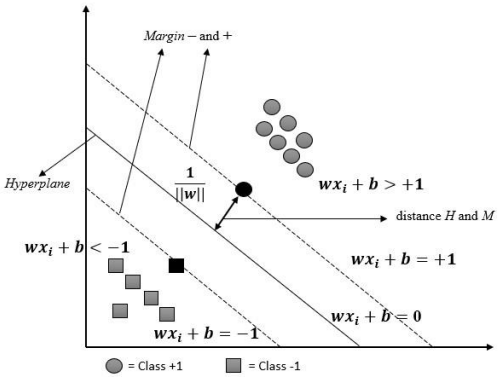

Figure 2 SVM's concept

The SVM concept based on Figure 2 is described as an attempt to find the best hyperplane that functions as a separator for two classes -1 and +1. The classification problem can be explained by trying to find the hyperplane that separates the two groups.

The best separating hyperplane between the two classes is found by measuring the hyperplane margins and finding the maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called a support vector. Trying to find the location of the hyperplane is the core of the learning process in SVM. The hyperplane equation assumes that both classes -1 and +1 can be perfectly separated by the d-dimensional hyperplane defined in equation (4).

$$\vec{w}\vec{x} + b = 0 \quad (4)$$

where:
$\vec{w}$ = Hyper field
$\vec{x}$ = To map each input vector into dimensional space
b = Bias

The pattern $\overrightarrow{x_i}$ which belongs to class -1 (negative sample), can be formulated as a *pattern* that satisfies inequality (5).

$$\vec{w}\vec{x} + b < -1 \quad (5)$$

While the pattern $\overrightarrow{x_i}$ which belongs to class +1 (positive sample), is formulated by inequality (6).

$$\vec{w}\vec{x} + b \geq 1 \quad (6)$$

The most significant margin can be found by maximizing the value of the distance between the hyperplane and its closest point, which is $\frac{1}{\|w\|}$ which is equivalent to minimizing$\|\vec{w}\|^2$. This can be formulated $\min_{\vec{w}} \tau(w)$ as a *Quadratic Programming* (QP) *problem*, namely finding the minimum point of equation (7) by paying attention to the constraints of equation (8).

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \quad (7)$$

$$y_i(\overrightarrow{x_i}.\vec{w} + b) - 1 \geq 0, \forall i \quad (8)$$

where:
y_i = Correspondence (submission of intent) class
$\overrightarrow{x_i}$ = Input vector

This problem can be solved by various computational techniques, including the *Lagrange Multiplier*.

$$L(\vec{w}, b, \alpha) = \frac{1}{2}|\vec{w}|^2 - \sum_{i=1}^{l}\alpha_i \left(y_i\left((\overrightarrow{x_i}.\vec{w} + b) - 1\right)\right)$$

with $(i = 1,2,\dots,l)$ (9)

α_i is a *Lagrange multipliers*, which are zero or positive $\alpha \geq 0$. The optimal value of equation (6) can be calculated by minimizing *L* respect to $\vec{w}$ and *b*, and maximizing *L* on α_i . By paying attention to the property that at the optimal point of the gradient *L*=0, equation (10) can be modified as maximization of problems containing only α_i, such as equation (11).

Maximum:
$$\sum_{i=l}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=l}^{l}\alpha_i\,\alpha_j y_i y_j \overrightarrow{x_i}\overrightarrow{x_j} \quad (10)$$

Subject:
$$\sum_{i=l}^{l}\alpha_i\, y_i = 0 \text{ ,with } \alpha_i \geq 0 (i = 1,2,\dots,l) \quad (11)$$

From the calculation results, it is obtained that α_i is primarily positive. Data that is correlated with positive α_i is what is referred to as a *support vector*.

2.7 Model Goodness Measures

The size of the model's goodness can be seen using the *Accuracy* value obtained from the *confusion matrix*. The *confusion matrix* is a table consisting of the number of rows of test data that are predicted to be correct and incorrect by a classification model (Lee *et al*., 2022). Illustration of the confusion matrix can be seen in Table 1

Table 1 Illustration of the *Confusion Matrix*

Actual Value	Predictive Value	
	Negative	Positive
Negative	True Negative (*TN*)	False Positive (*FP*)
Positive	False Negative (*FN*)	True Positive (*TP*)

The confusion matrix results used in this study are the value of *Accuracy* and F1-*score*. *Accuracy* is the number of correct predictions on all predicted data. *Accuracy* is a method of measuring the goodness of a model that is commonly used in classification modeling. F1-*score* is a comparison of the weighted average *Precision* and *Recall*. F1-*score* is a good measure of 7 good models used on unbalanced data. Calculation of the value of the *confusion matrix* can be done with the following formula.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (11)$$

$$Precission = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

$$F1 - score = 2\,x\,\frac{(Recall\ x\ Precision)}{(Recall+Prescision)} \quad (14)$$

where:
TP : The number of correct positive class predictions
TN : The number of positive class predictions is wrong

FP	:	The number of positive classes predicted incorrectly
FN	:	The number of classes other than positive is predicted to be true
Recall	:	The ratio of correctly positive predictions compared to the overall correct positive data
Precision	:	The ratio of positive correct predictions compared to the overall positive predicted outcome

Another model, apart from using the accuracy value, can also be seen using the AUC (area under curve). AUC is a good evaluation model for unbalanced data. AUC is the area under the ROC (receiver operating characteristic) curve ranging from zero to one (Wang *et al.*, 2017). A list of AUC range values and their classification levels, according to Gorunescu (2011), can be seen in Table 2.

Table 2 AUC range values and classification levels according to Gorunescu (2011)

AUC value	Classification level
0,91-1,00	Excellent
0,81-0,90	Good
0,71-0,80	Fair
0,61-0,70	Poor
0,50-0,60	Failure

3. Methodology

3.1 Research Procedure

The research procedure in this research includes the stages of literature study, data collection by scraping, dataset review, data labeling, text preprocessing series, weighting with TF-IDF, a combination of SMOTE and Tomek Links, training data and test data, modeling with the SVM method, evaluation with a confusion matrix. The research procedure can be seen in Figure 3.

3.2 Research Procedure

The data collection process is carried out using scrapping techniques on the online target market *https://shopee.co.id/*. The data taken in this study is part of consumer reviews and ratings, the data is the result of reviews and ratings in the form of stars with a scale of one to five stars from consumers. The scraping process can be seen in Figure 4.

The steps for the scraping process in figure 4 begin with taking the url of the destination store and the url of the API at *"https://shopee.co.id/"*, after the url is linked the data absorption process through scraping runs then the data that originally has the *.json* extension is parsed into the extension CSV and database processing in CSV form as the final result of scraping consumer review data.

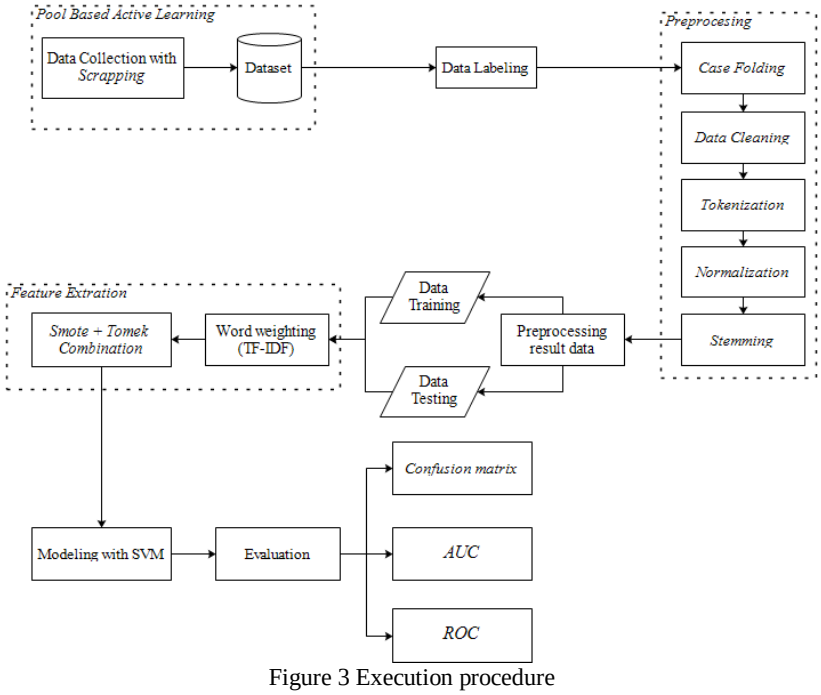

Figure 3 Execution procedure

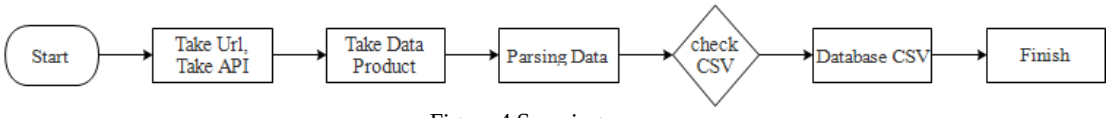

Figure 4 Scraping process

3.3 Data Labelling

The dataset taken in this study is still in a state of raw data that does not yet have a target class, so it is necessary to label the dataset. The process in this labelling stage is carried out with rating features and expert judgment or linguist validators to determine the pattern of review data on positive or negative sentiment. This classification is carried out on reviews that have a rating of 4 and 5 as positive sentiment and ratings of 3 to 1 as negative sentiment.

3.4 Preprocessing

Text *preprocessing* refers to the process of converting human language text into machine-interpretable text that can be used for further processing for predictive modeling tasks. The dataset uses Indonesian language review data contained in the review text. This process is very important to clean and prepare text data for the purposes of this research. The following are the stages of the preprocessing process. The editorial result of the original *preprocessing* is in Indonesian, which is translated into English. The following results of *preprocessing* data can be seen in Table 3.

4. Result And Discussion

4.1 Result

The method used as an evaluation in this study is the Confusion Matrix and the ROC AUC score. The test used 8628 data samples with a comparison of the results of data *Accuracy* using the SVM method without SMOTE Tomek links and the SVM method with a combination of SMOTE and Tomek Links.

4.2 *SMOTE and Tomek Links combination*

In Figure 5, the data distribution experiences data imbalance, unbalanced data can produce biased results towards the majority class and result in poor classification performance.

Table 3 *Preprocessing* results

No	Initial Sentence	Output Sentence	Label
1	For cool material but a bit thin. It's worth it at that price. For the size, the bottom is a bit short, overall is good ðŸ‘ðŸ‘" ðŸ‘•ðŸ‘•"	'for', 'material', 'cool', 'rather', 'thin' 'worth', 'it', 'with', 'price', 'that much'. 'for', 'size', 'under', 'rather', 'short', 'overall', 'good' '	Positive
2	In accordance with the description, bought at a flash sale price, thank you very much seller and mas courier y gercepðŸ™ðŸ™",",5	'according', 'description', 'buy' hrg 'flash', 'sale', 'thank you', 'many', 'seller'	Positive
3	"The goods are very good. The quality never disappoints. It is highly recommended to buy, so that your outfit is up to date. Maylova â¤ï¸",",5	'item', 'very', 'good' 'quality', 'does not disappoint', 'recommended', 'make', 'buy', 'biar', 'outfit, 'contemporary'	Positive
.	.	.	.
.	.	.	.
.	.	.	.
8626	This is so small, mah size M where is the babyterry material huui ðŸ‘Ž,3	'small', 'very', 'sizem', 'material', 'babyrry'	Negative
8627	They only use photos to attract buyers	'they', 'only', 'use', 'photo', 'for', 'interesting', 'buyer'	Negative
8628	The material doesn't match expectations, it's very disappointing that it doesn't match what is drawn, the material is thin but not too thick, sorry I'll just give 3 stars hehe:)",3	'Material', 'not', 'according', 'expectation', 'very,'disappointing', 'no', 'according to','picture', 'material', 'thin', 'sorry','love' 'star'	Negative

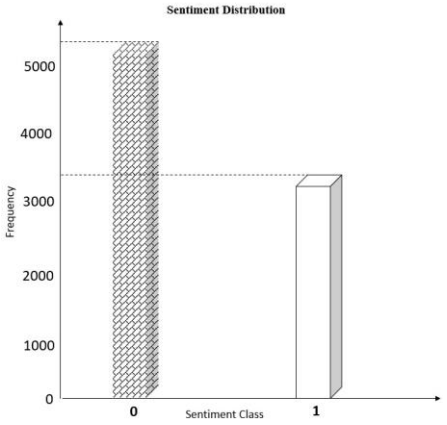

Figure 5 Data distribution without SMOTE-Tomek Links

Figure 6 shows the confusion matrix results used to calculate the *Accuracy*, *Precision*, *Recall*, and F1-*score* results of the SVM method without using a combination of SMOTE and Tomek links.

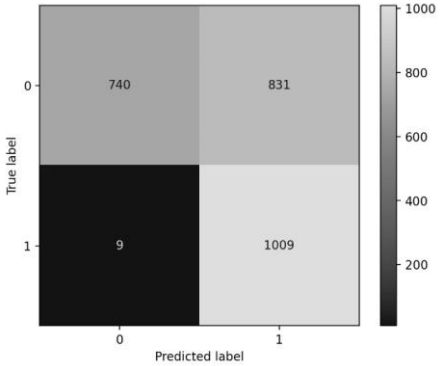

Figure 6 Confusion matrix without SMOTE- Tomek Links

The result of the confusion matrix in Figure 6 is that the *TP* value indicates the acquisition of the number of correct positive class predictions of 1009, the *TN* value indicates the acquisition of the number of incorrect prediction values of 740, the *FP* value indicates the acquisition of the number of positive class predicted incorrectly of 831, the *FN* value indicates the number of classes other than the positively predicted correct value of 9.

Evaluation with other methods using the ROC AUC curve score with the Support Vector Machine method without and using a combination of the SMOTE and Tomek Links methods. The results of the curve can be seen in Figure 7.

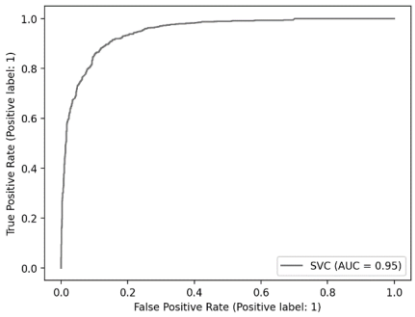

Figure 7 ROC AUC Without the Combination of SMOTE and Tomek Links

The results of the curve in Figure 7 show the ROC curve showing a curve skewed towards number 1 and an AUC score of 0.95 with SVM without using a combination of SMOTE and Tomek Links.

4.3 SVM performance with SMOTE and Tomek combination

In Figure 8, the data distribution uses a combination of SMOTE and Tomek Links which produces synthetic data for the minority class on SMOTE and Tomek Links to remove data identified as Tomek links from the majority class, that is, sample data from the majority class closest to the minority class data.

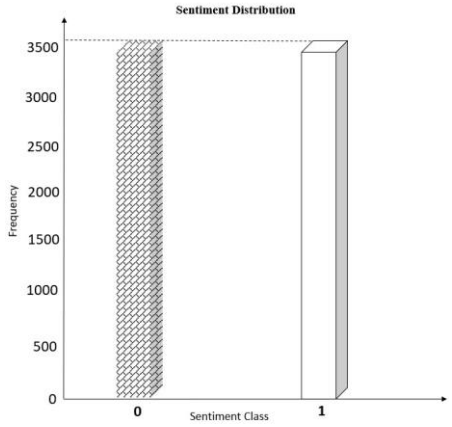

Figure 8 Data distribution with SMOTE-Tomek Links

Figure 8 shows the confusion matrix results used to calculate the *Accuracy*, *Precision*, *Recall*, and F1-*score* using the SVM method with a combination of SMOTE and Tomek links.

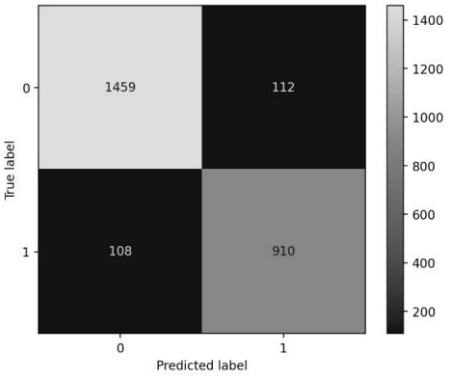

Figure 9 Confusion matrix with SMOTE-Tomek Links

The result of the confusion matrix in Figure 9 is that the *TP* value indicates the acquisition of the number of correct positive class predictions of 901, the *TN* value indicates the acquisition of the number of incorrect prediction values of 1459, the *FP* value indicates the acquisition of the number of positive class predicted incorrectly of 112, the *FN* value indicates the number of classes other than positively predicted value of 108.

Evaluation with other methods using the ROC AUC curve score with the SVM method using a combination of SMOTE and Tomek Links. The results of the curve can be seen in Figure 10.

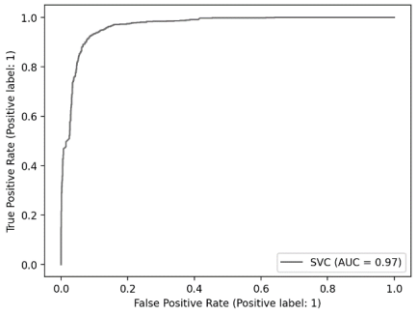

Figure 10 ROC AUC with SMOTE and Tomek Links Combination

The ROC AUC curve results show that the ROC curve results show a skewed curve towards number 1 and an AUC score of 0.97 with the SVM method using a combination of SMOTE and Tomek Links.

4.2 Discussion

In the initial process, the dataset was collected from the results of scrapping reviews *https://shopee.co.id/*, then the data went through preprocessing stages. At the feature extraction learning stage, namely TF-IDF and the combination of SMOTE with Tomek Links, then the data is divided into training data and test data. The next stage is entering the learning process of sentiment analysis classification using the Support Vector Machine algorithm method and as an evaluation using the calculation results of *Accuracy*, *Precision*, *Recall*, and F1-*score* from the confusion matrix. Table 4 shows the results of *Accuracy*, *Precision*, *Recall*, and F1-*score.*

Table 4 Comparison of *Confusion Matrix* Results

No	Item	Without SMOTE-Tomek Links	With SMOTE-Tomek Links
1	*Accuracy*	0,68	0,92
2	*Precision*	0,55	0,89
3	*Recall*	0,99	0,89
4	F1-*score*	0,71	0,89

In Table 4, based on Figures 6 and 9, it can be seen that the performance of SVM with the SMOTE and Tomek Links combination is better than SVM without SMOTE and Tomek Links. This happens because the number of class input data is not balanced, this can be seen in Figures 5 and 8, which are shown in the distribution of data.

ROC AUC curve displays performance information of the classification algorithm in graphical form. The ROC AUC curve is made based on the values obtained in the Confusion Matrix calculations. The results of comparing the AUC ROC scores with the Support Vector Machine method without and using a combination of the SMOTE and Tomek Links methods can be seen in Table 5.

Table 5 Comparison of AUC ROC scores

No	Model	Score of ROC AUC
1	Without SMOTE -Tomek Links	0,95
2	Using SMOTE-Tomek Links	0,97

The AUC ROC scores shown in Table 5 without SMOTE-Tomek Links is 0.95 while using SMOTE-Tomek Links, it has a score of 0.97. The increase in the AUC ROC score in the combination of SMOTE and Tomek Links was due to balancing the majority and minority class data with the combination of SMOTE and Tomek Links. The AUC range and classification level, according to Gorunescu (2011), 0.91 – 1.00 with a ROC curve inclined towards number 1, shows excellent results.

he results of sentiment analysis research using the SVM method with a combination of SMOTE and Tomek Links yield an accuracy value of 0.92 and an AUC score of 0.97 with a curve towards number 1. These results are better than sentiment analysis research without using data balancing techniques conducted by Singla *et al.*, (2017), who applied three methods of Naïve Bayes, Support Vector Machine (SVM), and Decision Tree with the highest accuracy score of 81.77. The following research was carried out by Arivoli and Sonali (2021), who examined sentiment analysis using the Polarity film review dataset taken from the Twitter API using the Naïve Bayes and Support Vector Machine classification methods. The results of the accuracy of the two scores of the Naïve Bayes algorithm are 76.67, and the accuracy score of the Support Vector Machine algorithm is 78.18.

Consumer review data has been validated by expert judgment or a linguist validator from the Department of Indonesian Literature, Faculty of Humanities, Sam Ratulangi University, so that the consumer review instrument can be used as calculation data in the evaluation process. Evaluation using the Confusion Matrix method resulted in better SVM performance with a combination of SMOTE and Tomek Links compared to SVM without a combination of SMOTE and Tomek Links.

5. Conclusion

Consumer reviews of products marketed through online marketplaces have been identified using sentiment analysis using the Support Vector Machine method and a combination of the Synthetic Minority Oversampling Technique (SMOTE) with Tomek Links. The results of 8628 clean data that have been classified using the SVM method and combined with SMOTE and Tomek Links produce Accuracy values of 0.92, Precision 0.89, Recall 0.89, and F1-score 0.89 with a difference of 0.03 between the values Accuracy with Precision, Recall and F1-score values are better than the SVM method without the combination of SMOTE and Tomek Links. The ROC AUC results of the combination of SMOTE and Tomek Links show that the ROC curve is inclined to number 1, and the AUC score is 0.97, which indicates excellent results. This is because the SVM method with a combination of SMOTE and Tomek Links has succeeded in identifying by balancing the majority and minority data classes, thereby increasing the performance of the classification results.

References

Arivoli, Sonali., 2021. Sentiment Analysis Using Support Vector Machine Based On Feature Selection and Semantic Analysis, International Research Journal of Computer Science 7 (8), 209-214.

Batista, G. E., Bazzan, A. L., Monard, M. C., 2003. Balancing training data for automated annotation of keywords: a case study. In WOB 10-18.

Bin Alias, M.S., Ibrahim, N.B., Zin, Z.B., 2021. Improved sampling data workflow using smtmk to increase the classification accuracy of imbalanced dataset. European Journal of Molecular and Clinical Medicine 8 (2), 91–99.

Baek, Y., Yun, U., Kim, H., Nam, H., Lee, G., Yoon, E., Vo, B., Lin, J.C.W., 2020. Erasable pattern mining based on tree structures with damped window over data streams. Engineering Applications of Artificial Intelligence 94, 103735.

Borg, A., Boldt, M., 2020. Using VADER sentiment and SVM for predicting customer response sentiment. Expert Systems with Applications 162113746.

Chapelle, G., Eymeoud, J.B., 2022. Can big data increase our knowledge of local rental markets? A dataset on the rental sector in France. PLoS ONE 17, 1–21.

Cortes C., Vapnik V., 1995, Support-vector networks. Machine Learn 20 (3), 273–297.

Elhassan AT., Aljourf M., Al-Mohanna F., Shoukri M., 2016. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (RUS) as a data reduction method, Global Journal of Technology & Optimization, 1-11.

Gorunescu, F., 2011. Data Mining Concepts, Models and Techniques. Springer, Berlin.

Hadwan, M., Al-Sarem, M., Saeed, F., Al-Hagery, M.A., 2022. An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and SMOTE technique. Applied Sciences 12 (11), 5547.

Hunt, I., 2021. In-sample tests of predictability are superior to pseudo-out-of-sample tests, even when data mining. International Journal of Forecasting 150156–154.

Lee, L.-H., Chen, C.-H., Chang, W.-C., Lee, P.-L., Shyu, K.-K., Chen, M.-H., Hsu, J.-W., Bai, Y.-M., Su, T.-P., Tu, P.-C., 2022. Evaluating the performance of machine learning models for automatic diagnosis of patients with schizophrenia based on a single site dataset of 440 participants. European Psychiatry 65 (1).

Li, X., Wu, C., Mai, F., 2019. The effect of online reviews on product sales: A joint sentiment-topic analysis. Information and Management 56 (2), 172–184.

Mustopa, A., Hermanto, Anna, Pratama, E.B., Hendini, A., Risdiansyah, D., 2020. Analysis of user reviews for the pedulilindungi application on google play using the support vector machine and naive bayes algorithm based on particle swarm optimization. 2020 5th International Conference on Informatics and Computing, ICIC 2.

Obiedat, R., Qaddoura, R., Al-Zoubi, A.M., Al-Qaisi, L., Harfoushi, O., Alrefai, M., Faris, H., 2022. Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution. IEEE Access 1022260–22273.

Sain, H., Purnami, S.W., 2015. Combine sampling support vector machine for imbalanced data classification. Procedia Computer Science 7259–66.

Sasada, T., Liu, Z., Baba, T., Hatano, K., Kimura, Y., 2020. A resampling method for imbalanced datasets considering noise and overlap. Procedia Computer Science 176420–429.

Singla, Z., Randhawa, S., Jain, S., 2017. Sentiment analysis on product reviews using machine learning techniques. 2017 International Conference on Intelligent Computing and Control (I2C2).

Swana, E.F., Doorsamy, W., Bokoro, P., 2022. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors 22 (9).

Wang, C., Chen, J., Chen, X., 2017. Pricing and order decisions with option con- tracts in the presence of customer returns. International Journal of Production Economics 193(1), 422-436.

Wang, Q., Zhang, W., Li, J., Mai, F., Ma, Z., 2022. Effect of online review sentiment on product sales: the moderating role of review credibility perception. Computers in Human Behavior 133, 107272.