❒ 2060

Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia

Dinar Ajeng Kristiyanti, Samuel Ady Sanjaya, Vinsencius Christio Tjokro, Jason Suhali
Department of Information System, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

Article Info	ABSTRACT
	Global recession news dominates social media, particularly in Indonesia, with social news platforms on Twitter generating public responses and re-tweetings on the issue. Mining these opinions from Twitter using a sentiment analysis approach yields invaluable insights. The research stages included data collection, pre-processing, data labeling using the lexical-based method like valence aware dictionary and sentiment reasoner (VADER) and TextBlob, sampling techniques using synthetic minority oversampling technique (SMOTE) and random over sampling (ROS) before and after splitting data, and modeling using machine learning such as support vector machines (SVM), k-nearest neighbour (KNN), naive Bayes, and model evaluation. The problem is that almost 300,000 data collected from NodeXL are unbalanced. The findings show that models with balanced datasets show better model evaluation results. The sampling technique was carried out before and after splitting the data. The model evaluation results show that the Bernoulli-naive Bayes algorithm, with the VADER labeling technique, and the SMOTE sampling technique after splitting data, obtains the best accuracy of 84%, and using the ROS technique obtains an accuracy of 81%. On the other hand, with the SMOTE and ROS technique before splitting data on the SVM algorithm, it gets the best accuracy of 93% from before if only using SVM only reached 84%. *This is an open access article under the [CC BY-SA](#) license.*

Corresponding Author:

Dinar Ajeng Kristiyanti
Department of Information System, Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
Email: dinar.kristiyanti@umn.ac.id

1. INTRODUCTION

The World Bank has predicted that there will be a massive global economic recession in 2023. It is supported by an aggressive increase in central bank interest rates, such as the US central bank reaching 3-3.25%, which is a sign that the recession in 2023 is not just a rumor [1]. A recession is a country's economic condition that has decreased and hurt the continuity of a country. This recession occurred in all parts of the world, where almost all experienced drastic economic paralysis. All food prices and other community needs experienced a significant increase in price increases. It makes many countries fall into a prolonged economic crisis [2]. The economy would enter a global recession zone in all countries, including Indonesia are also affected by the recession. This global recession has made several countries experience economic problems that are difficult to overcome, such as inflation, the energy crisis, the insufficient supply of food resources, and the financial crisis [3]. Based on data from the International Monetary Fund or IMF, it is said that there has been a consistent slowdown from 2021 to early 2023 [4]. The 2023 recession has

become a hot topic that everyone is talking about, including on social news platforms on social media like Twitter [5].

All news media are competing to provide information regarding the global recession [6]. News media are an important source of information for the public during epidemic crises, serving as interactive community bulletin boards and global or regional monitors [7]. With the prevalence of social media, news media organizations have used social media to reach and engage audiences during crises [8]. Social media is a new phenomenon in the world of information and communication technology (ICT) because it can attract internet users to interact with each other [9]. In Indonesia, there are 160 million people actively using social media [10]. Social media, such as Twitter, has changed it is function to become an adequate means for people to express their opinions on various matters [11]. The topic of conversation for cyberspace citizens often becomes a trending topic when many people express opinions about the topic [12]. It is in line with the number of official news portals with blue ticks on Twitter which also report on the issue of this recession. It then provides various public opinions on Twitter by re-tweeting, giving rise to various polemics, fear, and public anxiety over the existing issues, but also positively emerging various opinions as a preventive strategy to deal with them.

Processing tweets and retweets on Twitter will take a long time if each tweet's meaning is analyzed individually. Conversely, a little analysis will speed up processing time, but the information obtained becomes less relevant. Therefore, appropriate analysis techniques are needed to analyze many texts relatively quickly [13]. Sentiment analysis is a technique that is executed automatically to obtain personal information to understand sentiment from text data sources [14]. Sentiment analysis is a technique to extract text data to obtain information about positive, neutral, or negative sentiments [15]. Sentiment analysis also shows sadness, joy, or anger [16]. The text referred to in sentiment analysis can be in the form of news, product, and community reviews on social media. Many popular machine learning algorithms, such as the naïve Bayes (NB), support vector machine (SVM), k-nearest neighbors (KNN), decision tree (DT), linear regression (LR), artificial neural network (ANN), and random forest (RF) algorithm are used for sentiment analysis. The NB algorithm can classify data well based on probabilistic reasoning [17]. The SVM algorithm makes predictions in classification and regression cases [18]. Many researchers are looking for a way to reduce the complexity of KNN, which can be divided into three general methods, namely reducing the dimensions of the vector text, reducing the number of training samples, and speeding up the process of finding the k closest neighbors [19].

Many studies related to sentiment analysis have been carried out, such as predicting presidential candidates [20], evaluating product reviews [21], and much more. One study obtained an accuracy of 79.45% using KNN, which was higher than DT, NB, and RF algorithms [22]. Several studies on sentiment analysis that compared the SVM algorithm and other classification algorithms obtained an F1 score of 73.69% [23]. Based on different studies, SVM produces better values than NB, KNN, LR, and ANN [24]. In another study, the accuracy with the NB algorithm was 80.6%, and the SVM accuracy was 79.3% when using seven-fold cross-validation [25].

Several techniques are used to label a text in sentiment analysis based on it is polarity, namely by using a labeling technique. There are many approaches to labeling sentiments in text, namely using a lexicon-based method, deep learning, or manually. This research focuses on automatic labeling techniques using the valence aware dictionary and sentiment reasoner (VADER) and TextBlob algorithms. VADER is a sentiment analysis method used to determine a class in a sentence by giving a label based on the lexicon library. The labels provided can be negative, neutral, and positive labels [26]. Apart from VADER, other automatic labeling techniques are commonly used in sentiment analysis, such as TextBlob. TextBlob is an automated labeling technique in Python programming used to weight words given negative and positive labels based on the lexicon. TextBlob is generally used on textual data types for complex operations. One of the studies used VADER and TextBlob. The accuracy produced by VADER is 30%, TextBlob produces an accuracy of 25%, and bidirectional encoder representations from transformers (BERT) produces an accuracy of 51.36%. It proves that BERT produces higher accuracy than VADER and TextBlob in that study [27]. Another study showed that using the bag of words (BoW) feature on TextBlob resulted in higher accuracy than VADER, namely 86% and 82%, respectively [28]. However, other studies show that VADER's accuracy is 79% and TextBlob's is 73% [29]. Other research also indicates that BERT has an accuracy of 94%, which is higher than VADER 61% and TextBlob 62% [30]. Meanwhile, other studies compare IndoBERT, SVM, and NB. The F1 score accuracy obtained from this study is IndoBERT 84%, SVM 70%, and NB 83% [31]. Even though IndoBERT has pretty good accuracy compared to SVM and NB, the execution time for IndoBERT is around 5 hours, while for SVM and NB, it is around 15 minutes.

Imbalance data is the most common thing in sentiment analysis [32]. Data imbalance is a situation where the data ratio is not proportional or unbalanced, causing the performance of the model application to be ineffective [33]. Therefore, an oversampling technique is needed to balance the class distribution. Oversampling is a technique for balancing data distribution by increasing the distribution of low data to the same as other high data distributions [34]. There are several oversampling techniques, such as synthetic

minority oversampling technique (SMOTE) and random over sampling (ROS). SMOTE is the most popular oversampling technique, synthesising a new sample from a lower class to balance the distribution of the existing classes [35]. ROS is an oversampling technique that balances data distribution by randomly taking data until it meets the data needed to balance it [36]. The research uses the SMOTE oversampling technique against the NB and SVM algorithms. The results of the F1 score accuracy of NB and SVM that do not use SMOTE are 35.9% and 56.6%, while NB and SVM that use SMOTE produce higher accuracy, namely 91.4% and 91.9% [34]. Then other studies compare the SMOTE, adaptive synthetic (ADASYN), ROS, and data augmentation oversampling techniques. SMOTE has an accuracy value of 95.94%, train 99.86%, and validation of 96.41% [36]. Contributions in this study are: i) applying sentiment analysis related to the recession, ii) the dataset used in this research is Indonesian language tweets data, especially on news portal accounts, iii) comparing popular classification algorithms, namely NB, SVM, and KNN, in classifying sentiment, iv) comparing labeling techniques such as VADER and TextBlob related to the recession in sentiment analysis, v) overcoming data imbalance using oversampling techniques, such as SMOTE and ROS.

2. THE COMPREHENSIVE THEORETICAL BASIS

Mustaqim *et al.* [22] discusses the government's response to forest fires that occur, using datasets taken from Twitter, with a sentiment analysis approach using the VADER labeling technique and the KNN algorithm. Before the data were analyzed, preprocessing was carried out, which included case folding, cleaning, lemmatization, removing stopwords, and stemming. The accuracy results obtained are pretty high, namely 79.45% using the KNN algorithm compared to other algorithms, such as DT, NB, and RF. Stock market predictions using microblogging sentiment analysis and machine learning have also been carried out [25]. The algorithms used are KNN, SVM, LR, NB, DT, RF, and multilayer perceptron (MLP). The novelty of this research is that it integrates several sentiment analysis and machine learning techniques. It also suppresses taking additional features from social media, such as public sentiment, to increase the accuracy of stock predictions. Labeling techniques such as VADER and Textblob are also used. The result is SVM classification model combined with the VADER labeling technique, F1 score of 76.3%, and an area under curve (AUC) of 67%.

User response sentiment analysis has also been carried out on accessibility in mobile applications [28]. The classification algorithms used are LR, SVM, extra tree (ET), gaussian naïve Bayes (GNB), gradient boosting (GB), and AdaBoost. It is also used labeling techniques such as TextBlob and VADER. The result is that the TextBlob algorithm has a more significant percentage than VADER. However, for the GNB and GB algorithms, it has an increase of 3%, namely 68% and 84%. Another sentiment analysis study discusses the sentiment analysis of Islamophobia during the church attack on social media Twitter [34]. The labeling technique used is VADER. The classification algorithm used is NB and SVM. Then the SMOTE technique is also used to balance imbalanced data. The result is that the data performed by the SMOTE technique has a higher accuracy than the raw data. For NB and SVM, accuracy before SMOTE is 73% and 81%. After SMOTE, the NB and SVM increased to 90.8% and 91.3%.

Hate speech detection has also been done using machine learning classification [37]. This study uses 4,002 Twitter datasets related to politics, religion, ethnicity, or certain races in Indonesia. The classification algorithms used are NB, MLP, AdaBoost classifier, DT, and SVM. Because the data used does not have a balanced distribution, the oversampling technique used is SMOTE. The highest accuracy results obtained for the SMOTE classification algorithm are multinomial naïve Bayes of 73.2%.

Decision support system for heart disease prediction based upon machine learning [38]. SVM, NB, LR, RF, and AdaBoost classifiers are the classification algorithms used. Because the data used does not have an even distribution, the SMOTE oversampling technique is used. The highest accuracy obtained by using oversampling is NB of 85.07%. Other research discusses malware detection in android applications with a machine-learning approach on imbalanced datasets. The classification algorithm used is KNN, SVM, and iterative dichotomiser, where the algorithm is used as a detection model. To generalize the distribution of data, the SMOTE oversampling technique is used. The result is that the KNN algorithm has the highest accuracy, precision, recall, f-measure, and Matthews correlation coefficient (MCC), respectively, 98.69%, 97.89%, 99.49%, 98.69%, and 97.39% [39]. A dataset of student performance is used to make predictions with classification techniques supported by the ensemble voting method [40]. The classification algorithm used is NB, KNN, conjunctive rules, and Hoeffding tree. Oversampling is also done using the SMOTE method to equalize the data distribution. The accuracy results were quite significant, with NB of 95.5% getting the highest accuracy.

Classify customer messages on e-commerce sites using supervised learning. DT, NB, SVM, and LR are the classification algorithms used. The dataset used has an uneven distribution. Therefore, an oversampling method, such as the ROS, is applied. The SVM algorithm is obtained, which has the most

significant level of accuracy, amounting to 78.5%. Previous literature studies on previous sentiment analysis that studied oversampling, labeling, or classification approaches are summarized in Table 1.

Table 1. The previous sentiment analysis studies about oversampling, labeling, or classification approaches

Related work	Oversampling technique	Labeling technique	Classifier method	Best result
[22]	-	VADER	KNN, DT, NB, RF	VADER+KNN, acc 79.45%
[25]	-	VADER, TextBlob	KNN, SVM, LR, NB, DT, RF, MLP	VADER+SVM, F1-score 76.3%, AUC 67%
[28]	-	VADER, TextBlob	LR, SVM, ET, GNB, GB, AdaBoost	TextBlob (BoW feature), acc 86%
[34]	SMOTE	VADER	NB, SVM	SMOTE+VADER+SVM, acc 91.39%
[37]	SMOTE	-	NB, MLP, AdaBoost, DT, SVM	SMOTE+NB, acc 73.2%
[38]	SMOTE	-	SVM, NB, LR, RF, AdaBoost	SMOTE+NB, acc 85.07%
[39]	SMOTE	-	KNN, SVM, ID	SMOTE+KNN, acc 98.69%
[40]	SMOTE	-	NB, KNN, CR, HT	SMOTE+NB, acc 95.5%
[41]	ROS	-	DT, NB, SVM, LR	ROS+SVM, acc 78.5%

3. METHOD

3.1. Data collection

In this study, the dataset was taken from social media Twitter which contains comments on Indonesia's recession from January 2023 to May 2023. These comments have various meanings. There have been some comments condemning the recession by blaming the government for comments that provide solutions to the recession that our nation will face. The amount of data used is 300,000, which still needs cleaning. Therefore, the data will be further processed at the pre-processing stage before being used in the machine learning model. In data cleaning, duplicate data is removed. After cleaning, data clean yields 38,000 records. Cleaned data will then be labeled based on the sentiment contained therein. Labels are divided into 3 categories, namely positive, negative, and neutral. If the comment is considered positive, it will be denoted by the number "1". If negative, it will be denoted by "0". If neutral, it will be denoted by the number "2".

This study uses 2 architectural models, namely splitting-oversampling and oversampling-splitting, it can be seen in Figure 1. The stage starts with collecting data from Twitter and then preprocessing the data. Data preprocessing includes several processes, including removing URLs and mentions, removing punctuations, tokenization, removing stop words, stemming, removing words with length < 3, joining words back, and translating using Google Translate. Then, the data labeling process was carried out using the VADER and Textblob libraries. In the next stage, 2 decisions will be used, which include the method approach, namely splitting-oversampling and oversampling-splitting. The difference is in the data processing stage. In oversampling-splitting stage, oversampling is performed first using the SMOTE and ROS methods. Then, data splitting was carried out with a proportion of 80% for the training and 20% for the testing classes.

In splitting-oversampling method, the splitting is done first by 80% for training and 20% for testing. In the exercise, oversampling is carried out to balance data distribution using the SMOTE and ROS methods. Ultimately, it will be included in the machine learning model using the NB, SVM, and KNN algorithms. The modeling results will be made into an evaluation model to find the most significant accuracy.

3.2. Data cleansing

The first stage in data pre-processing is data cleansing. Data cleansing is the process of modifying or deleting data that is considered inaccurate, duplicate, incomplete, malformed, or damaged in the owned dataset. Data cleansing makes it possible to delete data that is not needed so that the data is clean and can improve accuracy when entered into the algorithm. This study uses six stages of data cleansing, including: i) remove URLs and mentions, ii) punctuation removal, iii) tokenization, iv) stop words removal, v) stemming, and vi) remove irrelevant data. Recombine the words that all stages of data cleansing have processed into the data list into a sentence string that represents the tweets textually. So, this string data can be used for sentiment analysis.

3.2.1. Remove URLs and mentions

Data taken from Twitter generally cannot be separated from URLs and mentions that do not provide relevant information for text analysis [42]. If URLs and mentions are not cleaned up, it will be difficult for the algorithm to analyze or determine the sentiment of data. In this study, removing URLs and mentions from text uses the re library, a regular expression operation.

3.2.2. Punctuation removal

This removes punctuation marks because punctuation marks such as periods, commas, and question marks, often do not have an essential meaning in text analysis. By removing punctuation marks, it can

simplify the text and focus on the main words only [42]. In deleting punctuation, use the re library with the re.sub() method to set all characters that are not letters, numbers, or spaces. Then the tweet's text is changed to lowercase with method.lower(). Changing text to lowercase is not without reason. It is to help in text consistency and avoid differences in meaning due to differences in capitalization.
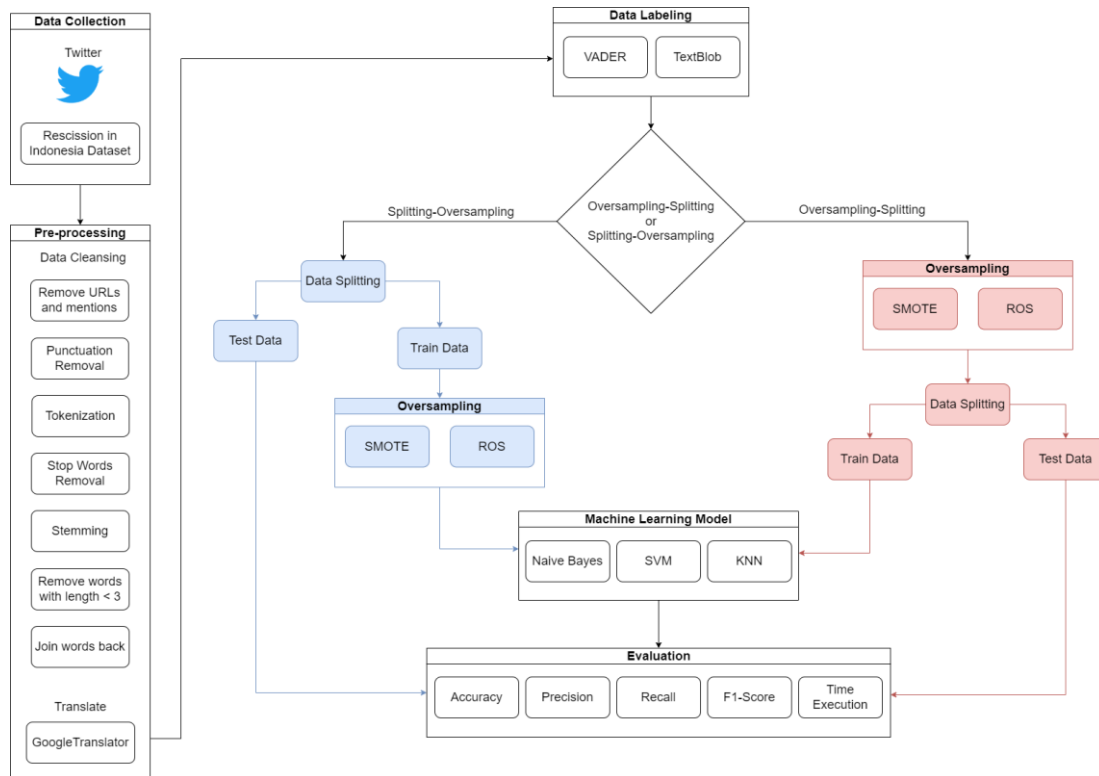

Figure 1. Proposed architecture oversampling-splitting and splitting oversampling

3.2.3. Tokenization

Tokenization breaks down a sentence into words based on spaces and punctuation, which later weigh a word based on sentiment [43]. Tokenization is the process of simplifying text by breaking words down into tokens-units that are considered semantically helpful. Depending on the scale, tokenization divides sentences into one full text (sentence tokenization) or words into one sentence (word tokenization). Tokenization uses the nltk library (natural language toolkit) with the word_tokenize() method.

3.2.4. Stop words removal

Stopword removal is part of the text preprocessing stage, which aims to remove irrelevant words in a sentence based on the stopword list. Delete words that are common and have no critical meaning, for example, "the", "and", "is", which often appear but do not provide meaningful information. So that the text being analyzed can focus on essential words and reduce the size of the vocabulary, deletion is performed using a list comprehension to produce a list of words not included in the stop_words list [42].

3.2.5. Stemming

It turns words into basic words by removing the inflexion of words [44]. It is done to reduce the variation of words that have the same root so that it can be considered as one entity, and it can reduce the complexity of the text and produce a simpler representation. Stemming is done using the nltk library with a stemmer.

3.2.6. Remove irrelevant data

Elimination of very short words such as "a", "an", and "in" which do not provide much information or are not meaningful, especially in text analysis [45]. Data is filtered of less than three characters to delete these words. So, that only long words are counted.

3.3. Data translation

The data taken is data in Indonesian. It is necessary to translate it into English. It works so that it can be used by algorithms that only accept English-language data in providing data labels. This study uses the Google Translate API to translate data from Indonesian to English.

3.4. Data labeling

The next stage is to label the data. The labeling process adds target attributes. It is necessary because the analysis was carried out in this study using a machine-learning approach based on supervised learning. Generally, two methods of labeling data are manual, semi-automatic, and automatic. This research uses automatic labeling, including VADER and TextBlob.

3.4.1. Valence aware dictionary and sentiment reasoner

The VADER determines the polarity of positive, negative, or neutral text data into labels. VADER's ability to recognize the emotional intensity and negative words in the text makes the accuracy results relatively high [35]. However, it should be noted that when using VADER, the text data it can receive is English. It is why the data was translated into English in the previous stages.

3.4.2. TextBlob

Similar to VADER, TextBlob can only analyze English text. What distinguishes VADER and TextBlob is understanding sentences contextually using linguistic rules. It allows TextBlob to recognize the context and nuances contained in the text by considering the use of words, grammar, and sentence construction to get positive, negative, or neutral sentiments to be included in the label [27].

3.5. Data splitting

Data splitting into training and testing data is needed to validate at the end. In this study, data splitting was carried out using a ratio of 75:25 for training data and testing data [46]. Where the training data is used to be trained through oversampling and machine learning models, then the model's results based on the training data will be validated by data testing so that the validation results can measure how effective the proposed architecture is in overcoming the sentiment analysis problem.

3.6. Oversampling technique

Data collected, for example, from social media, may need to be more balanced or is commonly referred to as imbalanced data [34]. It is called imbalanced data if several samples are significantly unbalanced, causing a majority and minority class. If the imbalanced data is corrected for analysis, it will ensure the results are accurate. To crush imbalanced data, there is a technique called oversampling [35].

3.6.1. Synthetic minority oversampling technique

One method commonly used to deal with oversampling is SMOTE. Using SMOTE will create a new synthetic sample in the minority class by combining existing samples [39]. The new sample results are obtained from the differences between the selected features and their neighbors based on random sample selection from the minority class and looking for the closest neighbors. So that SMOTE can distribute data evenly.

3.6.2. Random over sampling

The difference between SMOTE and ROS lies in how the new data is generated. ROS works by duplicating or repeating an existing sample from the minority class randomly [41]. Then replicate it until the number is balanced with the majority class.

3.7. Machine learning model

The final step is to classify data based on sentiment using a machine-learning approach to training data. This study uses several popular classification algorithms for classification in sentiment analysis. The machine learning algorithms include NB, SVM, and KNN.

3.7.1. Naïve Bayes

Using the NB algorithm, data can be classified into different sentiment categories using positive, negative, and neutral probability calculations, as shown in (1). NB is suitable for use with high-dimensional data because it is fast and simple [47]. The likelihood of event A happening given the occurrence of event B (conditional probability) is denoted as P(A|B). P(B|A) represents the probability of event B occurring given the evidence of event A. P(A) signifies the probability of event A happening, while P(B) represents the probability of event B occurring.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad (1)$$

3.7.2. Support vector machine

Data classification using the SVM algorithm can separate data into classes based on text representation into positive, negative, or neutral categories. SVM is also famous for being able to work well on complex, non-linear, and overfitting data. It is because SVM can learn patterns and relationships between features and related sentiments and maximize the distance between samples from different classes to find the best hyperplane [37], as shown in (2):

$$w \cdot x + b = 0 \qquad (2)$$

3.7.3. K-nearest neighbors

Data classification using the KNN algorithm can classify text based on most classes from its nearest neighbors in the feature space to determine the most common sentiment category in sentiment analysis. To use KNN, a specified K value is required, as shown in (3). The K value is the number of nearest neighbors used for the classification [47]. To calculate the distance between two points in the KNN algorithm, the Euclidean distance method is used, which can be used in 1-dimensional space, 2-dimensional space, or multi-dimensional space. 1-dimensional space means that the distance calculation uses only one independent variable, 2-dimensional space means that there are two independent variables, and multi-dimensional space means that there are more than two variables.

$$d(x,y) = (\sum_{i=1}^{m}|xi - y_i|^r)^{\frac{1}{r}} \qquad (3)$$

3.8. Evaluation

After all the processes, models have been created using the training data, and the testing data will be used to test model performance or validation. Evaluation aims to gauge, appraise, and judge the model's effectiveness. The assessment conducted in this research employed the confusion matrix to evaluate accuracy, precision, recall, F1 score, and execution time. Validation is executed to confirm the model's reliability using available data. The validation performed in this study utilizes cross-validation.

4. RESULTS AND DISCUSSION

This section will further explain the results of model classification, oversampling techniques, and labeling on sentiment analysis of the recession in Indonesia. The results of the classification of machine learning models include accuracy, precision, recall, F1 score, and time execution. In the final section, a comparison will be made between oversampling-splitting and splitting-oversampling. The parameters used for comparison are oversampling techniques, such as VADER and Textblob. Oversampling techniques, such as SMOTE and ROS. Classification models used, such as KNN, NB, and SVM.

4.1. Experimental comparison classification result based on labeling technique

Table 2 compares sentiment labeling based on two labeling techniques, VADER and TextBlob. With VADER, a significant portion of the data is detected as negative sentiment, leading to data imbalance. On the other hand, when using TextBlob, the sentiment distribution is more balanced between positive and neutral sentiments, but the occurrence of negative sentiment is relatively low compared to both positive and neutral sentiments.

Two bar charts illustrate the division of sentiments more clearly for a clearer picture. Figure 2(a) illustrates the comparison of labeling techniques using VADER. Figure 2(b) illustrates the labeling technique using TextBlob to classify sentiments as positive, negative, or neutral based on their respective accuracies.

Table 2. Comparison of sentiment label results based on labeling technique

Labeling technique	Sentiment label		
	Positive	Negative	Neutral
VADER	8718	2592	26782
TextBlob	15186	16480	6426
VADER-SMOTE	26782	26782	26782
VADER-ROS	26782	26782	26782
TextBlob-SMOTE	16150	16150	16150
TextBlob-ROS	16150	16150	16150

The sentiment labeling using VADER and TextBlob was both unbalanced, with a substantial portion of the data identified as a negative sentiment, as in Figure 2(a), and a significant portion of the data was identified as positive and neutral sentiment in Figure 2(b). This imbalance rendered the data less relevant when utilized in machine learning algorithms. As a result, implementing the oversampling method was necessary to balance the number of samples between the majority and minority classes. By employing oversampling on the positive and neutral sentiment data, a more balanced dataset was achieved, ensuring better representation and improved performance in model training. Oversampling was used to balance the data, increasing accuracy when classifying using machine learning algorithms. The oversampling technique employed included SMOTE as shown in Figures 3(a) and 4(a), and ROS as shown in Figures 3(b) and 4(b).

(a)   (b)
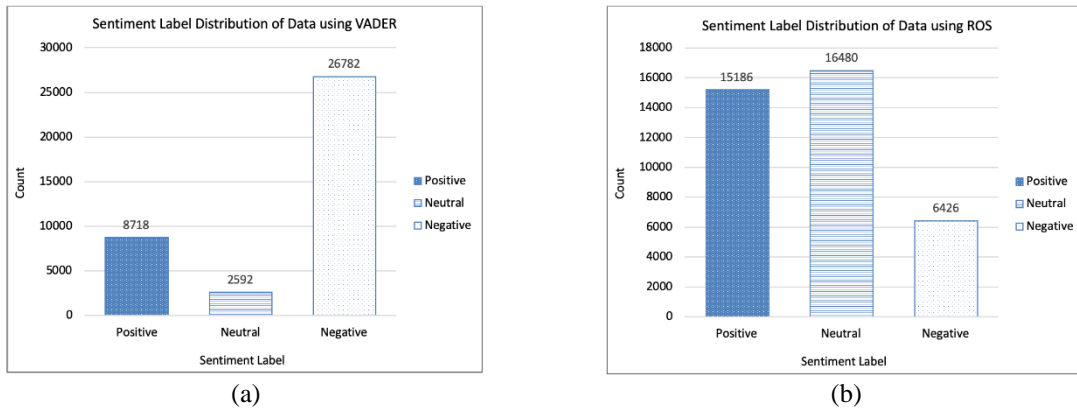
Figure 2. Sentiment label distribution of data using (a) VADER and (b) ROS

(a)   (b)
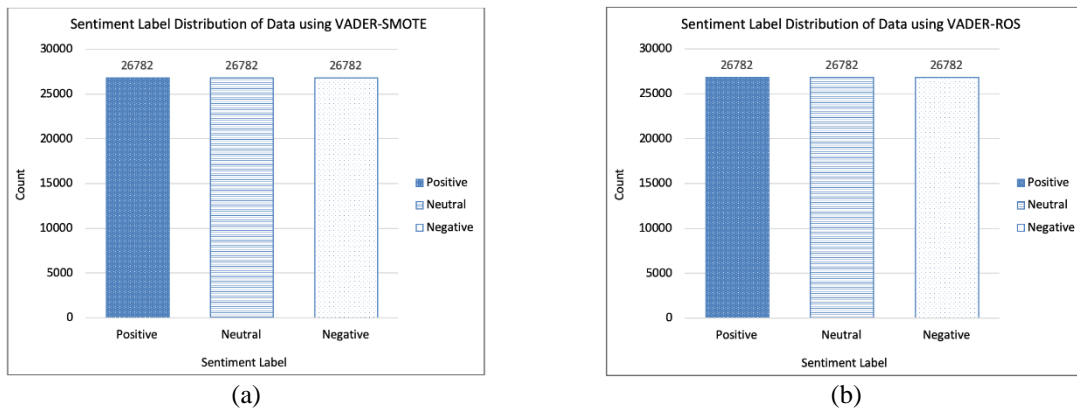
Figure 3. Sentiment label distribution of data using (a) VADER-SMOTE and (b) VADER-ROS

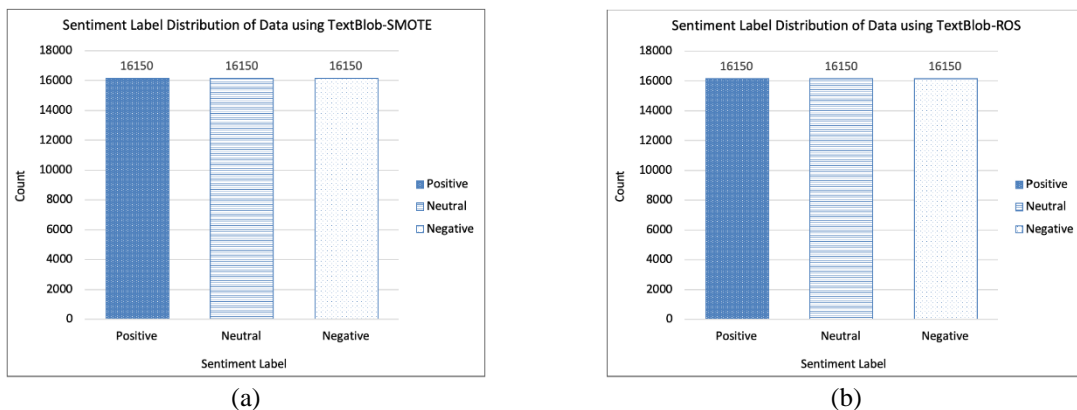(a)   (b)

Figure 4. Sentiment label distribution of data using (a) TextBlob-SMOTE, (b) TextBlob-ROS

The use of oversampling with the SMOTE and ROS method for the VADER labeling technique divides the data into equal parts based on sentiment, and all label sentiments have the same number of 26,782 data. So, using SMOTE and ROS produce the same distribution when labeling with VADER. This is the same when using oversampling with the VADER and ROS methods for the TextBlob labeling technique. SMOTE and ROS share the same amount of data for all sentiment labels, namely 16150. Thus, using SMOTE and ROS results in equal division when labeling with TextBlob.

4.2. Experimental result of splitting-oversampling method

Table 3 show the result of the splitting-oversampling method with the classification of the working duration of each model, which is made differently in the labeling, oversampling, and classification model method. In Table 3, the splitting process with ratio 75:25 is carried out first and then oversampling on the classification model. In the VADER library without oversampling, the SVM algorithm gets the highest accuracy rate of 84% and an F1 score of 62%. In the VADER library with the SMOTE oversampling method, the NB algorithm gets the highest accuracy rate of 84% and an F1 score of 83%. In the VADER library with the ROS oversampling method, the NB algorithm gets the highest accuracy rate of 81% and an F1 score of 81%. Then the results can be determined that the SMOTE oversampling model has the highest accuracy and F1 score in the VADER library.

In the TextBlob library without oversampling, the SVM algorithm gets the highest accuracy rate of 84% and F1 score of 84%. In the Textblob library with the SMOTE oversampling method, the NB algorithm gets the highest accuracy rate of 78% and an F1 score of 75%. In the TextBlob library with the ROS oversampling method, the SVM algorithm gets the highest accuracy rate of 76% and an F1 score of 76%. Then the results can be determined that the model without the oversampling method has the highest accuracy and F1 score in the TextBlob library. In the splitting-oversampling architectural model, the highest level of model accuracy is using the Textblob labeling method and without the oversampling process.

Table 3. Confusion matrix from experimental result of splitting-oversampling method

Labeling technique	Over sampling technique	Classifier method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time execution (ms)
VADER	-	KNN	78	65	55	58	40
VADER	-	Naïve Bayes	76	53	48	50	15,000
VADER	-	SVM	84	83	84	62	68,000
VADER	SMOTE	KNN	46	59	55	43	130
VADER	SMOTE	Naïve Bayes	84	84	84	83	16,500
VADER	SMOTE	SVM	79	80	79	79	215,000
VADER	ROS	KNN	49	60	54	44	50
VADER	ROS	Naïve Bayes	81	81	81	81	11,000
VADER	ROS	SVM	79	80	79	80	178,000
TextBlob	-	KNN	62	74	55	58	30
TextBlob	-	Naïve Bayes	69	67	65	66	18,000
TextBlob	-	SVM	84	84	84	84	110,000
TextBlob	SMOTE	KNN	59	65	53	54	60
TextBlob	SMOTE	Naïve Bayes	67	63	63	63	14,000
TextBlob	SMOTE	SVM	78	75	75	75	188,000
TextBlob	ROS	KNN	64	66	60	60	30
TextBlob	ROS	Naïve Bayes	68	65	66	65	18,000
TextBlob	ROS	SVM	76	73	74	76	218,000

4.3. Experimental result of oversampling-splitting method

Table 4 show the result of oversampling-splitting method with the classification of the working duration of each model, which is made differently in the method of labelling, oversampling, and the classification model. In Table 4, the oversampling process is carried out first. Then splitting is carried out to proceed to make the classification model. In the VADER library without oversampling, the SVM algorithm gets the highest accuracy rate of 84% and F1 score of 62%. In the VADER library with the SMOTE oversampling method, the SVM algorithm gets the highest accuracy rate of 93% and F1 score of 93%. In the VADER library with the ROS method, the SVM algorithm gets the highest accuracy rate of 93% and an F1 score of 93%. In the SMOTE and ROS libraries, the SVM algorithm has the same level of accuracy and F1 score, but the precision and recall are higher in the SMOTE method. Then the results can be determined that the SMOTE oversampling and SVM classification models are more suitable for use.

In the TextBlob library without oversampling, the SVM algorithm gets the highest accuracy rate of 84% and an F1 score of 84%. In the TextBlob library with the SMOTE oversampling method, the SVM algorithm gets the highest accuracy rate of 85% and F1 score of 86%. In the Textblob library with the ROS

oversampling method, the SVM algorithm gets the highest accuracy rate of 85% and an F1 score of 86%. In the SMOTE and ROS libraries, the SVM algorithm has the same level of accuracy and F1 score. Therefore, the results can be determined that the model with the SMOTE and ROS methods has the highest accuracy and F1 score in the TextBlob library.

Table 4. Confusion matrix from experimental result of oversampling-splitting method

Labeling Technique	Over Sampling Technique	Classifier Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Time Execution (ms)
VADER	-	KNN	78	65	55	58	40
VADER	-	Naïve Bayes	76	53	48	50	15,000
VADER	-	SVM	84	83	84	62	68,000
VADER	SMOTE	KNN	84	83	84	62	100
VADER	SMOTE	Naïve Bayes	84	84	84	83	13,500
VADER	SMOTE	SVM	93	94	94	93	212,000
VADER	ROS	KNN	49	60	54	44	50
VADER	ROS	Naïve Bayes	81	81	81	81	11,000
VADER	ROS	SVM	93	93	93	93	229,000
TextBlob	-	KNN	62	74	55	58	30
TextBlob	-	Naïve Bayes	69	67	65	66	18,000
TextBlob	-	SVM	84	84	84	84	110,000
TextBlob	SMOTE	KNN	76	80	76	76	60
TextBlob	SMOTE	Naïve Bayes	73	74	73	73	260,000
TextBlob	SMOTE	SVM	85	86	86	86	200,000
TextBlob	ROS	KNN	77	81	77	76	30
TextBlob	ROS	Naïve Bayes	72	73	72	71	18,000
TextBlob	ROS	SVM	85	86	86	86	210,000

4.4. Discussion

This study focuses on sentiment analysis and examines the impact of oversampling techniques used to remove imbalanced data, often overlooked in sentiment analysis studies. The oversampling technique used in this study is SMOTE and ROS to balance the data set, ensuring consistent results in machine learning models. There are several points of interest if we compare this study with previous research on sensitivity analysis. For example, conducted a sensitivity analysis of government responses to forest fires using the VADER labeling technique and KNN algorithm [22]. Although the KNN algorithm achieved a reasonable accuracy of 79.45%, the study did not consider the data balance of techniques such as oversampling.

On the other hand, this research uses an oversampling method with two different methods, splitting-oversampling compared to oversampling-splitting, where the technique first runs oversampling, then splitting for higher accuracy. It is because the oversampling procedure increases the number of minority samples so that the results obtained when the data are balanced produce more accurate results during the experiment. Like preceding studies, we explored stock market predictions through sentiment microblogging analysis and system mastering using diverse category algorithms [25]. This look obtained an F1 score of 76.3% and an AUC of 67% using VADER labeling combined with SVM. While the oversampling method was not used, it proved critical in this study.

Another relevant research dealt with Islamophobia sentiment evaluation for the duration of a church assault on Twitter [34]. The study uses VADER for labeling and SMOTE for balancing records, resulting in much better accuracy. It is in step with the findings of this look which highlight the importance of using oversampling techniques, particularly in conducting sentiment evaluation. Furthermore, centered on detecting hate speech through diverse class algorithms, wherein SMOTE is likewise used for unequal records distribution [37]. The accuracy acquired is 73.2% using multinomial NB. Similarly, Rani *et al.* [38] explored coronary heart sickness detection with the system getting to know, and oversampling through SMOTE improved the accuracy of the NB classifier to 85.07%.

In the context of Android malware detection, confirmed the effectiveness of the SMOTE approach with the KNN algorithm, achieving a high accuracy of 98.69% [39]. Their findings corroborate the blessings of employing oversampling strategies to enhance class effects. Finally, centered on classifying client messages on e-trade websites through supervised mastering with the SVM algorithm. The research uses ROS to stability records and produces 78.5% accuracy [41]. Overall, this study gives treasured insights for enhancing sentiment analysis techniques within the destiny, emphasizing balancing facts and decisions for each splitting-oversampling and oversampling-splitting approach. By comparing the results with preceding paintings, it is far located that oversampling techniques can enhance accuracy and higher version performance, particularly in unbalanced information.

5.   CONCLUSION

This study focused on sentiment analysis related to the global recession using Indonesian-language tweets. The dataset consisted of 38,000 tweets collected from news portal accounts on Twitter. The data underwent several stages of data preprocessing, including data cleansing, translation from Indonesian to English, and labeling using automatic techniques such as VADER and TextBlob. The labeled data was split into training and testing sets in a 75:25 ratio. To address data imbalance, oversampling techniques, specifically SMOTE and ROS, were applied. The study compared popular classification algorithms in sentiment classification, namely NB, SVM, and KNN. It also compared labeling techniques such as VADER and TextBlob. Additionally, the study explored the impact of oversampling techniques and the order of splitting and oversampling in the classification process. Various parameters were considered, with a focus on accuracy as well as other relevant metrics. In the splitting-oversampling approach, the highest accuracy was achieved using VADER-SMOTE-Naïve Bayes (BernoulliNB) with 84% accuracy, 84% precision, 84% recall, 83% F1 score, and an execution time of 16,500 ms. TextBlob-SVM achieved 84% accuracy, 84% precision, 84% recall, 84% F1 score, and an execution time of 110,000 ms. In the oversampling-splitting approach, the highest accuracy was obtained with VADER-SMOTE-SVM at 93% accuracy, 94% precision, 94% recall, 93% F1 score, and an execution time of 212,000 ms. TextBlob-SMOTE-SVM achieved 85% accuracy, 86% precision, 86% recall, 86% F1 score, and an execution time of 200,000 ms. Overall, this research contributes to sentiment analysis by applying it to global recession using Indonesian-language tweets. It compares classification algorithms, labeling techniques, oversampling techniques, and splitting and oversampling sequences. The findings highlight the importance of choosing the right method based on the specific context and requirements of the analysis, particularly in addressing data imbalances. The oversampling and labeling technique has successfully dealt with unbalanced data in sentiment analysis. Future research will compare other oversampling techniques such as BorderLine SMOTE, KMeans SMOTE, SVM SMOTE, ADASYN, and SMOTE-nominal and continuous (NC). Future research can also add other supervised classification methods or deep learning methods to improve further sentiment analysis capabilities, such as the accuracy and performance of sentiment analysis in various domains.

ACKNOWLEDGEMENTS

REFERENCES

[1]   J. D. Guenette, M. A. Kose, and N. Sugawara, "Is a global recession imminent?," *SSRN Electronic Journal*, pp. 1-47, 2022, doi: 10.2139/ssrn.4223901.
[2]   "Global economies 'out of sync' - HSBC asset management," *Funds Europe*, 2023. Accessed: Jul. 25, 2023. [Online]. Available: https://www.funds-europe.com/global-economies-out-of-sync-hsbc-asset-management/
[3]   "How should Indonesia navigate 2023's economic challenges?," *PwC Indonesia*, 2023. Accessed: Jul. 25, 2023. [Online]. Available: https://www.pwc.com/id/en/media-centre/press-release/2023/english/how-should-indonesia-navigate-2023-s-economic-challenges.html
[4]   D. Aprian, "Threat of recession 2023, President Jokowi: Indonesia's economy grows 5.44 percent in quarter II," *Voice of Indonesia*, 2022. Accessed: May. 20, 2023. [Online]. Available: https://voi.id/fr/economie/219830
[5]   J. Josephs, "Global recession warning as World Bank cuts economic forecast," *BBC News*, 2023. Accessed: May. 20, 2023. [Online]. Available: https://www.bbc.com/news/business-64213830
[6]   J. Strömbäck *et al.*, "News media trust and its impact on media use: toward a framework for future research," *Annals of the International Communication Association*, vol. 44, no. 2, pp. 139–156, Apr. 2020, doi: 10.1080/23808985.2020.1755338.
[7]   E. S. Al-Sheikh and M. H. A. Hasanat, "Social media mining for assessing brand popularity," *International Journal of Data Warehousing and Mining*, vol. 14, no. 1, pp. 40–59, 2018, doi: 10.4018/IJDWM.2018010103.
[8]   S. Kemp, "Digital 2022: Global overview report," *Data Portal*, 2022. Accessed: Feb. 02, 2023. [Online]. Available: https://wearesocial.com/sg/blog/2022/01/digital-2022-another-year-of-bumper-growth/
[9]   N. Azzouza, K. Akli-Astouati, and R. Ibrahim, "Twitterbert: framework for twitter sentiment analysis based on pre-trained language model representations," *Advances in Intelligent Systems and Computing*, vol. 1073, pp. 428–437, 2020, doi: 10.1007/978-3-030-33582-3_41.
[10]   A. L. Alten, G. Gadre, S. Kulkarni, and C. S. Wu, "Analyzing happiness index based on geographical locations," *2019 2nd International Conference on Artificial Intelligence and Big Data, ICAIBD 2019*, pp. 45–51, 2019, doi: 10.1109/ICAIBD.2019.8837010.
[11]   N. Berry, F. Lobban, M. Belousov, R. Emsley, G. Nenadic, and S. Bucci, "#WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems," *Journal of Medical Internet Research*, vol. 19, no. 4, pp. 1-13, 2017, doi: 10.2196/jmir.6173.
[12]   W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, "COVID-19 and the 5G conspiracy theory: social network analysis of twitter data," *Journal of Medical Internet Research*, vol. 22, no. 5, pp. 1-9, 2020, doi: 10.2196/19458.
[13]   D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM naïve bayes algorithm for sentiment analysis toward west java governor candidate period 2018-2023 based on public opinion on twitter," in *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, 2019, pp. 1-6, doi:

10.1109/CITSM.2018.8674352.

[14] H. Karayiğit, A. Akdagli, and Ç. İ. Aci, "Homophobic and hate speech detection using multilingual-BERT model on Turkish social media," *Information Technology and Control*, vol. 51, no. 2, pp. 356–375, 2022, doi: 10.5755/j01.itc.51.2.29988.

[15] G. Li, Q. S. Zheng, L. Zhang, S. Z. Guo, and L. Y. Niu, "Sentiment infomation based model for chinese text sentiment analysis," *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering, AUTEEE 2020*, pp. 366–371, Nov. 2020, doi: 10.1109/AUTEEE50969.2020.9315668.

[16] M. Abdullah, M. AlMasawa, I. Makki, M. Alsolmi, and S. Mahrous, "Emotions extraction from Arabic tweets," *International Journal of Computers and Applications*, vol. 42, no. 7, pp. 661–675, Oct. 2020, doi: 10.1080/1206212X.2018.1482395.

[17] Z. Guo, "Text classification based on naive bayes with adjusted weights via frequency ratio of feature words," *Proceedings-2021 International Conference on Computer Technology and Media Convergence Design, CTMCD 2021*, pp. 263–267, Apr. 2021, doi: 10.1109/CTMCD53128.2021.00063.

[18] F. H. Khan, U. Qamar, and S. Bashir, "Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach," *Soft Computing*, vol. 23, no. 14, pp. 5431–5442, Jul. 2019, doi: 10.1007/s00500-018-3187-9.

[19] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012, doi: 10.1016/j.eswa.2011.08.040.

[20] D. A. Kristiyanti, Normah, and A. H. Umam, "Prediction of Indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis," *Proceedings of 2019 5th International Conference on New Media Studies, CONMEDIA 2019*, pp. 36–42, 2019, doi: 10.1109/CONMEDIA46929.2019.8981823.

[21] D. A. Kristiyanti, D. A. Putri, E. Indrayuni, A. Nurhadi, and A. H. Umam, "E-wallet sentiment analysis using naïve bayes and support vector machine algorithm," *Journal of Physics: Conference Series*, vol. 1641, no. 1, pp. 1-6, 2020, doi: 10.1088/1742-6596/1641/1/012079.

[22] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," *Journal of Physics: Conference Series*, vol. 1567, no. 3, pp. 1-7, 2020, doi: 10.1088/1742-6596/1567/3/032024.

[23] O. Oyebode, R. Lomotey, and R. Orji, ""I Tried to Breastfeed but…": exploring factors influencing breastfeeding behaviours based on tweets using machine learning and thematic analysis," *IEEE Access*, vol. 9, pp. 61074-61089, 2021, doi: 10.1109/ACCESS.2021.3073079.

[24] N. Pombo, M. Rodrigues, Z. Babic, M. Punceva, and N. Garcia, "Computerised sentiment analysis on social networks. Two case studies: FIFA World Cup 2018 and Cristiano Ronaldo joining Juventus," *Trends and Applications in Information Systems and Technologies*, Cham, Switzerland: Springer Nature, pp. 126–140, 2021, doi: 10.1007/978-3-030-72651-5_13.

[25] P. Koukaras, C. Nousi, and C. Tjortjis, "Stock market prediction using microblogging sentiment analysis and machine learning," *Telecom*, vol. 3, no. 2, pp. 358–378, 2022, doi: 10.3390/telecom3020019.

[26] M. Shahzad and H. Alhoori, "Public reaction to scientific research via twitter sentiment prediction," *Journal of Data and Information Science*, vol. 7, no. 1, pp. 97–124, Feb. 2022, doi: 10.2478/jdis-2022-0003.

[27] K. Sharma and R. Bhalla, "Decision support machine-a hybrid model for sentiment analysis of news headlines of stock market," *International Journal of Electrical and Computer Engineering Systems*, vol. 13, no. 9, pp. 791-798, 2022, doi: 10.32985/ijeces.13.9.7.

[28] W. Aljedaani, F. Rustam, S. Ludi, A. Ouni, and M. W. Mkaouer, "Learning sentiment analysis for accessibility user reviews," *Proceedings-2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW 2021*. pp. 239–246, 2021, doi: 10.1109/ASEW52652.2021.00053.

[29] W. N. S. W. Min and N. Z. Zulkarnain, "Comparative evaluation of lexicons in performing sentiment analysis," *Journal of Advanced Computing Technology and Application*, vol. 2, no. 1, pp. 1–8, 2020.

[30] M. Alshehri, L. Alrajhi, A. Alamri, and A. I. Cristea, "MOOCSent: a sentiment predictor for massive open online courses," in *29th International Conference on Information Systems Development (ISD2021)*, pp. 1-12, 2021.

[31] Fransiscus and A. S. Girsang, "Sentiment analysis of COVID-19 public activity restriction (PPKM) impact using BERT method," *International Journal of Engineering Trends and Technology*, vol. 70, no. 12, pp. 281–288, 2022, doi: 10.14445/22315381/IJETT-V70I12P226.

[32] K. Ghosh, A. Banerjee, S. Chatterjee, and S. Sen, "Imbalanced twitter sentiment analysis using minority oversampling," in *2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019*, 2019, doi: 10.1109/ICAwST.2019.8923218.

[33] E. Kurniawan, F. Nhita, A. Aditsania, and D. Saepudin, "C5.0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in bandung regency," in *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, 2019, vol. 4, pp. 1–5, doi: 10.1109/ICoICT.2019.8835324.

[34] W. Gata and A. Bayhaqy, "Analysis sentiment about islamophobia when Christchurch attack on social media," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1819–1827, 2020, doi: 10.12928/TELKOMNIKA.V18I4.14179.

[35] M. I. Marwat *et al.*, "Sentiment analysis of product reviews to identify deceptive rating information in social media: a sentideceptive approach," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 3, pp. 830–860, 2022, doi: 10.3837/tiis.2022.03.005.

[36] A. Alhudhaif, B. Almaslukh, A. O. Aseeri, O. Guler, and K. Polat, "A novel nonlinear automated multi-class skin lesion detection system using soft-attention based convolutional neural networks," *Chaos, Solitons and Fractals*, vol. 170, May 2023, doi: 10.1016/j.chaos.2023.113409.

[37] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," *IOP Conference Series: Materials Science and Engineering*, vol. 830, no. 3, pp. 1-6, 2020, doi: 10.1088/1757-899X/830/3/032006.

[38] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, 2021, doi: 10.1007/s40860-021-00133-6.

[39] D. T. Dehkordy and A. Rasoolzadegan, "A new machine learning-based method for android malware detection on imbalanced dataset," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 24533–24554, 2021, doi: 10.1007/s11042-021-10647-z.

[40] M. Ashraf, M. Zaman, and M. Ahmed, "To ameliorate classification accuracy using ensemble vote approach and base classifiers," in *Emerging Technologies in Data Mining and Information Security*, Singapore: Springer , pp. 321-334, 2019, doi: 10.1007/978-981-13-1498-8_29.

[41] M. A. A. Sánchez and I. Galpin, "Classifying incoming customer messages for an e-commerce site using supervised learning," *Communications in Computer and Information Science*, vol. 1643, pp. 91–105, 2022, doi: 10.1007/978-3-031-19647-8_7.

[42] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions

for twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, pp. 298–310, 2018, doi: 10.1016/j.eswa.2018.06.022.

[43] T. Jo, *Text mining: concepts, implementation, and big data challenge*, Cham: Springer Nature, 2019, doi: 10.1007/978-3-319-91815-0.

[44] S. Wrycza and J. Maślankowski, "Social media users' opinions on remote work during the COVID-19 pandemic. Thematic and sentiment analysis," *Information Systems Management*, vol. 37, no. 4, pp. 288–297, 2020, doi: 10.1080/10580530.2020.1820631.

[45] D. Sharma and M. Sabharwal, "Sentiment analysis for social media using SVM classifier of machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, pp. 39–47, 2019, doi: 10.35940/ijitee.I1107.0789S419.

[46] M. Van M. Buladaco, "Sentiments analysis on public land transport infrastructure in davao region using machine learning algorithms," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 685–690, 2020, doi: 10.30534/ijatcse/2020/97912020.

[47] C. Kariya and P. Khodke, "Twitter sentiment analysis," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-3, doi: 10.1109/INCET49848.2020.9154143.

BIOGRAPHIES OF AUTHORS

Dinar Ajeng Kristiyanti 🆔 SC received her master's degree in Computer Science, and bachelor degree in Information System from Universitas Nusa Mandiri, Indonesia. Currently she is a Ph.D. candidate in Computer Science at IPB University, Indonesia. She is a lecturer at Department of Information System in Universitas Multimedia Nusantara, Indonesia. Her research interests include sentiment analysis, text mining, feature selection, optimization, data science, and machine learning. She has published over 13 papers in international journals and conferences. She can be contacted at email: dinar.kristiyanti@umn.ac.id.

Samuel Ady Sanjaya 🆔 SC holds a Master of Informatics degree from Institut Teknologi Bandung, Indonesia in 2019. He also received his Bachelor of Informatics from Atma Jaya Yogyakarta University, Indonesia in 2016, respectively. He is currently a lecturer at Department of Information Systems in Universitas Multimedia Nusantara, Indonesia. His research includes data analysis, machine learning, data mining, and social network analysis. He has published over 5 papers in international journals and conferences. He can be contacted at email: samuel.ady@umn.ac.id.

Vincencius Christiano Tjokro 🆔 SC currently enrolled as an undergraduate student, majoring in Information System at Universitas Multimedia Nusantara, Indonesia. His research areas of interest include artificial intelligent and MLOps. He can be contacted at email: vincencius@student.umn.ac.id.

Jason Suhali 🆔 SC holds an undergraduate Bachelor of Information System at Universitas Multimedia Nusantara, Indonesia. He is currently having internship at Sinarmas Land as Project Implementor Analyst. He is alumnus of Bangkit Academy 2023 led by Google, GoTo, and Traveloka in machine learning path. His research areas of interest include sentiment analysis, feature selection, and social network analysis. He can be contacted at email: jason.suhali@student.umn.ac.id.