Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal Vol. 6, No. 2, April 2024, pp: 148-156; eISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Manuscript received February 10, 2024; revised February 26, 2024; accepted March 8, 2024; date of publication April 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i2.381

How to cite: Noor Zalekha Al Habesyah, Rudy Herteno, Fatma Indriani, Irwan Budiman, and Dwi Kartini, Sentiment Analysis of TikTok Shop Closure in Indonesia on Twitter Using Supervised Machine Learning, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 2, pp. 148-156, April 2024

Sentiment Analysis of TikTok Shop Closure in Indonesia on Twitter Using Supervised Machine Learning

Noor Zalekha Al Habesyah, Rudy Herteno, Fatma Indriani, Irwan Budiman, and Dwi Kartini

Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia

Corresponding author: Rudy Herteno (email: rudi.herteno@ulm.ac.id)

ABSTRACT TikTok Shop is one of the features in TikTok application which facilitates users to buy and sell products. The integration of TikTok Shop with social media has provided new opportunities to reach customers and increase sales. However, the closure of TikTok Shop has caused controversy among the public. This study aims to analyze the views and responses of TikTok users in Indonesia to the closure of TikTok Shop. The dataset used was obtained from Twitter. The research methodology consists of labeling, oversampling, splitting, and machine learning, which includes SVM, Random Forest, Decision Tree, and Deep Learning (H2O). The contribution of this research enriches our understanding of the implementation of machine learning, especially in sentiment analysis of TikTok Shop closures. From the test results, it is known that Deep Learning (H2O) + SMOTE obtained AUC 0.900, without using SMOTE, AUC 0.867. SVM + SMOTE obtained AUC 0.885, without using SMOTE AUC 0.881. Random Forest + SMOTE obtained AUC 0.822, while without using SMOTE AUC 0.830. Decision Tree + SMOTE AUC 0.59; without SMOTE, AUC 0.646. Deep Learning (H2O) with SMOTE produces better performance compared to SVM, Random Forest, and Decision Tree. With an AUC of 0.900; it can be said that Deep Learning (H2O) has excellent performance for sentiment analysis of TikTok Shop closures. This research has significant implications for social electronic commerce due to its potential utilization by social media analysts.

INDEX TERMS Machine learning, sentiment analysis, deep learning, electronic commerce.

I. INTRODUCTION

Social commerce is a digital and efficient answer to the problems that traditional commerce faces[1]. TikTok Shop is a popular and currently trending e-commerce platform[1]. TikTok Shop is considered capable of competing with the Facebook marketplace and Instagram Shop, because TikTok Shop can offer attractive and intense benefits to users for purchasing and selling transactions[1], [2]. The closure of TikTok Shop in Indonesia has led to a variety of opinions from the public, particularly Twitter users. Some users are upset by the policy, while others support it.

Sentiment analysis is a way to explore opinions or texts that are mined from various social media platforms and use machine learning for the calculation process[3]. While machine learning has been widely used for sentiment analysis, there is an urgent need for an advanced approach[4]. Sentiment analysis is suitable to see how the public, especially Twitter users, responds to the policy[3], [5].

Sentiment analysis research was conducted by[6], using a dataset from Twitter as large as 17189, with Support Vector Machine (SVM) producing an accuracy rate of 0.89% and AUC of 0.8729. The study also used a Decision Tree which resulted in an accuracy rate of 0.81% and AUC of 0.8070. Another study, using a dataset from Kaggle with 1000 data points with Decision Tree and Random Forest resulted in AUCs of 0.704 and 0.732[7].

Compared to previous research, this study utilizes Twitter data that focuses on responses from users to the closure of the TikTok Shop in Indonesia. This research proposes a Deep Learning algorithm with a framework (H2O) that is already available in RapidMiner. RapidMiner is software for machine learning and data mining because it has flexible operators for data output and input in various file formats and contains more than 100 learning schemes for classification, regression, and clustering[8], [9], [10], [11].

This study aims to determine the performance of the model that has been built. The model is built with data splitting, data oversampling, and different classification algorithms. Each model will provide information about which combination has the highest performance. This model can be applied for sentiment analysis with other cases and datasets. The results of this study are expected to contribute as follows:

a. Provide an understanding of classification performance based on the number of datasets, division of training data, and testing data.
b. Provide knowledge about the effect of oversampling data with SMOTE.
c. Add insight into the most effective algorithm for sentiment analysis based on AUC value.

II. MATERIAL AND METHODS

The research process can be seen in FIGURE 1 which consists of data crawling, preprocessing, splitting, oversampling, cross-validation, and classification.

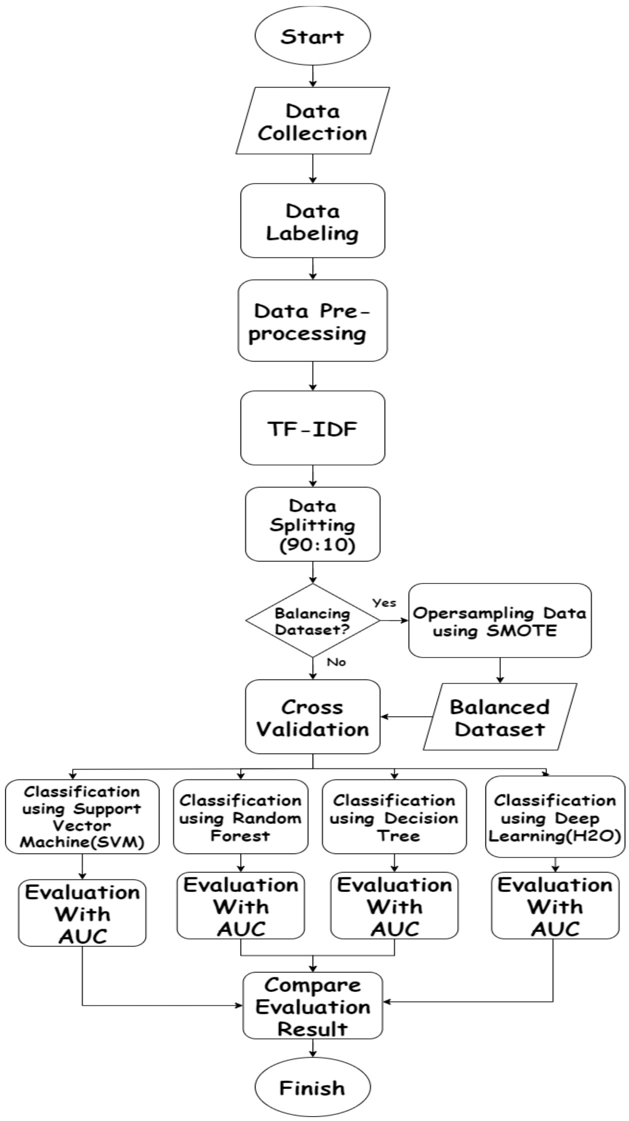

FIGURE 1. Research flow of SVM, Random Forest, Decision Tree, and Deep Learning (H2O) classification models.

A. DATA COLLECTION

Data collection is the process of obtaining tweet datasets by utilizing the API (Application Program Interface) from Twitter [12]. The tool used for this data collection is Google Collab with the Python programming language[6], with a set limit of 3000 data. Data collection was carried out from 12 November 2023 to 11 December 2023, during the TikTok Shop closure period in Indonesia, and 2903 data was obtained. Then the removal of the duplicate data was carried out using RapidMiner tools, resulting in 1624 pieces of data.

B. DATA LABELING

Data labeling is an important process to determine the sentiment of each tweet[13]. Data labeling in this research is done manually. Compared to machine labeling [13], this manual technique can provide accurate results without needing to see many examples, but it involves a lot of work by humans because it requires humans to read and analyze each dataset before labeling positive and negative[6], [7] .

In this research, 567 positive sentiments and 1057 negative sentiments were obtained (TABLE 1).

TABLE 1
Example of sentiment in tweet

Tweet	Sentimen
Because TikTok Shop is closed, many JNT employees are affected and have been laid off. Even many live hosts are also affected by the impact #trade minister https://t.co/5ONiBvtDnH”	Negative
I strongly support the banning of TikTok Shop so that the merchants in Cimol Gede Bage Market can prosper again... 👀 https://t.co/hyNfcirUmE	Positive

C. DATA PREPROCESSING

Data preprocessing is the process of correcting or removing damaged, miss formatted, or incomplete data sets [11], [14]. At this stage, RTs, URLs at the front and back of tweets, mentions, symbols, excessive spaces are removed, and converting numbers into text (TABLE 2) [4], [7].

TABLE 2
Result of first stage preprocessing

Input	Output
Because TikTok Shop is closed, many JNT employees are affected and have been laid off. Even many live hosts are also affected by the impact #trade minister https://t.co/5ONiBvtDnH”	Because TikTok Shop is closed, many JNT employees are affected and have been laid off. Even many live hosts are also affected by the impact
I strongly support the banning of TikTok Shop so that the merchants in Cimol Gede Bage Market can prosper again... 👀 https://t.co/hyNfcirUmE	I strongly support the banning of TikTok Shop so that the merchants in Cimol Gede Bage Market can prosper again 👀

The next preprocessing stage is carried out within the document from data, namely:

1) Transform Cases

In this process, all capital letters can be converted uniformly to lowercase characters, or vice versa [11]. In this research, words containing uppercase letters will be converted to lowercase (TABLE 3).

TABLE 3
Transform cases result

Input	Output
Because TikTok Shop is closed, many JNT employees are affected and have been laid off. Even many live hosts are also affected by the impact	because tiktok shop is closed, many jnt employees are affected and have been laid off. even many live hosts are also affected by the impact
I strongly support the banning of TikTok Shop so that the merchants in Cimol Gede Bage Market can prosper again... 👀	i strongly support the banning of tiktok shop so that the merchants in cimol gede bage market can prosper again... 👀

2) Tokenize
In this process, the string is divided into words, which are called tokens (TABLE 4).. There are two tokenization processes, the first is to remove emojis from the dataset, and the second is to split sentence datasets into words[7], [11], [15].

3) Filter Token by Length
This process involves determining the maximum and minimum token lengths (TABLE 5). In this study, the minimum token length is set to 3 letters, and the maximum token length is set to 25 letters[11].

TABLE 4
Tokenize and filter token result

Input	Output
because tiktok shop is closed, many jnt employees are affected and have been laid off. even many live hosts are also affected by the impact i strongly support the banning of tiktok shop so that the merchants in cimol gede bage market can prosper again... 👀	['because' 'tiktok' 'shop' 'closed' 'many' 'jnt' 'employees' 'are' 'affected' 'and' 'have' 'been' 'laid off' 'even' 'live' 'hosts' 'also' 'the' 'impact' 'stongly' 'support' 'banning' 'that' 'merchants' 'cimol' 'bage' 'gede' 'market' 'can' 'prosper' 'again']

4) Filter Stopword
This process eliminates non-essential or meaningless words such as "what", "how", "is", etc[7], [11]. The stopword dictionary utilized in this research is tala-stopword -Indonesia.

5) Stemming
Stemming is the process of removing affixes from words to produce their base form[7], [11]. The stemming dictionary utilized in this research is Sastrawi.

TABLE 5
Tokenize and filter token result

Input	Output
['because' 'tiktok' 'shop' 'closed' 'many' 'jnt' 'employees' 'are' 'affected' 'and' 'have' 'been' 'laid off' 'even' 'live' 'hosts' 'also' 'the' 'impact' 'stongly' 'support' 'banning' 'that' 'merchants' 'cimol' 'bage' 'gede' 'market' 'can' 'prosper' 'again']	['tiktok' 'shop' 'closed' 'jnt' 'employees' 'laid off' 'live' 'host' 'affected' 'support' 'banned' 'merchants' 'cimol' 'bage' 'gede' 'market' 'prosper']

Visualization of data from Negative class and Positive class (FIGURE 2 and FIGURE 3):

FIGURE 2. WordCloud Negative Class

FIGURE 3. WordCloud Positive Class

D. TF-IDF VECTORIZER

After the preprocessing is complete [6], [7], [11], the dataset will be transformed into vector form using RapidMiner with the "Process Document from Data" feature, selecting the TF-IDF vector creation parameter. Various algorithms can be applied to the dataset once it has been transformed into vector form[7]. All words are extracted and weighted according to their frequency of occurrence. TF-IDF calculation formula is given by Eq. (1) [7], [11] .

$$TF-IDF = \frac{p}{q} * \log\left(\frac{r}{s}\right) \tag{1}$$

where p indicates number of times appears in documents, q is total number terms in documents, r shows number of documents under consideration, and s indicates number of documents that contain the keyword.

E. DATA SPLITTING

Splitting data is the process of dividing a dataset into two parts, namely training data and testing data. Previous studies proposed training and testing data ratios of (70:30) or (80:20) to obtain landslide datasets, while for soil residual strength prediction, the ratios used were (70:30), (80:20), and (90:10)

[16]. In a previous study, increasing the Training Set Size (TSS) from 30% to 90% resulted in improved and more stable performance of the training data[17]. Therefore, a data split ratio of 90:10 was chosen for this research (TABLE 6).

TABLE 6
Data splitting result

	Positive	Negative	Total
Training Without Smote	510	951	1461
Testing Without Smote	57	106	163
Training With Smote	951	951	1902

F. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique aimed at augmenting minority class data to achieve balance within the dataset [18], [19], [20]. Two experiments were conducted during this phase. The first experiment did not use SMOTE, followed by the second experiment employing SMOTE. The objective here is to assess SMOTE's effectiveness in balancing data. Initially, based on the oversampling rate N, this process chooses N samples for every minority class case from K neighbouring samples belonging to minority classes. Following that, the SMOTE model generates N fresh instances according to Eq. (2) for a minority class. Finally, it integrates these newly created instances with the existing datasets.

$$x_{new} = x + rand(0,1) \times (y[i] - x) \tag{2}$$

where i is 1,2, … , N, rand(0,1) , random numbers between 0 and 1 are represented by the expression. x_{new} denotes the newly generated instance, while x represents an instance from the minority class. The term y[i] refers to the neighbour of x that is closest to i.[20], [21]

G. CROSS VALIDATION

Cross-validation is a technique widely used by researchers to evaluate the performance of classifiers, by randomly selecting between training and testing data samples and grouping data that have as much in common as the specified K value [22]. A common value for K is between 5 and 10 [23]. In this study, the value of K is determined to be 10.

H. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a machine learning method that constructs a hyperplane between positive and negative classes in a field, then uses it for classification [7], [24]. SVM are based on the definition of a hyperplane, which is given by Eq. (3) [25], [26].

$$\omega \cdot x + b = 0 \tag{3}$$

$$yi \cdot (xi + b) \geq 1 \tag{4}$$

the weight vector (ω) is orthogonal to the hyperplane, and the bias term (b) specifies the hyperplane's offset from the origin. SVMs minimise ½ |(|ω|)|^2 for all data points, subject to the limitations given in Eq. (3). The class labels are denoted by yi, and Xi represents the list of x. This research use dot product or linear kernel parameters. The SVM formula goes as follows Eq. (5):

$$K\,(xi,xj) = x.\,\mathrm{y} \tag{5}$$

In SVM, K is the kernel, while x and y are data points that make up a vector reflecting classification results [25], [27].

I. DECISION TREE

Decision Tree is a machine learning model for classification through inductive learning from known class data, presenting a structure similar to a tree with leaf nodes serving as class labels and internal nodes indicating prediction results [10] . Decision Tree takes a top-down approach to divide a data subset, and a variable and splitting boundary are selected at each stage of the process.Then, the dataset is repeatedly divided into pure subsets based on the impurity measure. The Decision Tree defines the goodness of split as the difference between the degree of impurity before and after division. Therefore, a greater purity in the divided data results indicates a higher goodness in the split. As a result, the data set is split through division boundary R with the highest goodness of split defined as (Eq. (6)):

$$\mathrm{G(T\,,R) = I(T\,) - I(T\,|R)} \tag{6}$$

where T is a set of the training example. G(T , R) indicates the goodness of split when the training set T is divided by R and I(T) and I(T |R) indicate the impurities before and after division based on the division boundary.Decision Tree applies its goodness of split criteria to each split point and evaluates the reduction in the impurity. Then, Decesion Tree selects the best split point of the variable in which the reduction in the impurity is the highest[28].

Decision Tree has impurity metrics that can be used to determine the splitting boundary. The impurity metrics are defined according to informatics and statistical approaches, such as the Information gain, Gini Index, gain ratio, distance measure[28], [29]. This research using Gini index because is one of the representative indices for measuring the impurity of data[28]. The following is the Decision Tree with the gini_index calculation (Eq. (7)):

$$G(n) = 1 - \sum_{i-1}^{c} p_i^2 \tag{7}$$

where G(n) is the Gini- impurity at node n and Pi specifies the proportion of observed class c at node n[7].

J. RANDOM FOREST

Random Forest is a machine learning algorithm belonging to the ensemble category consisting of numerous decision trees trained and each tree carrying bootstrapped, or commonly known as out-of-bag observations, samples for every observation[30]. For each instance, the Random Forest learning algorithm calculates an overall score by comparing actual observations against predictions derived from untrained subtree sets utilizing the specific observation, this overall score serves as a measure of Random Forest's

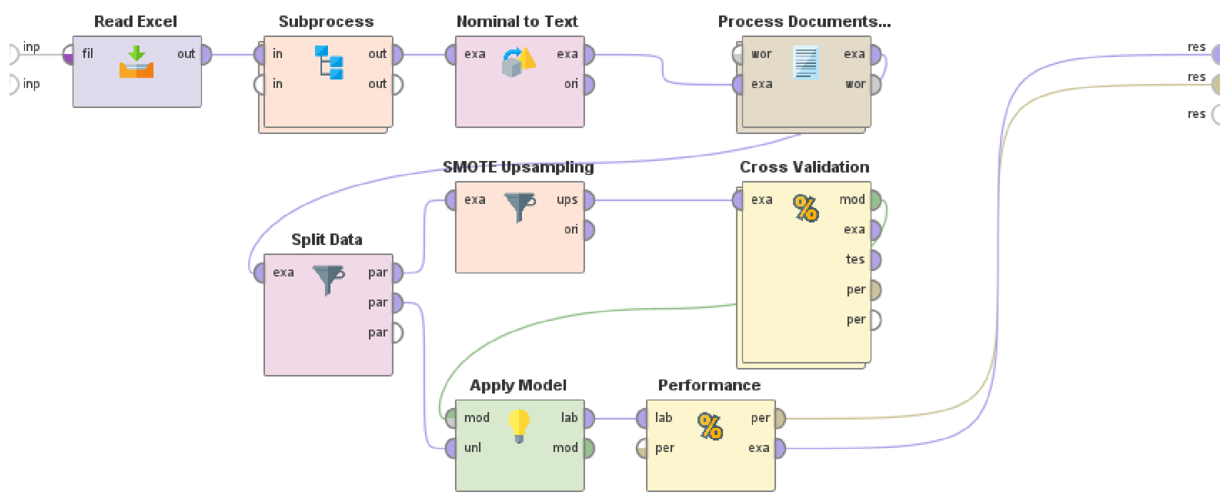

FIGURE 4. Research flow model data preprocessing, TF-IDF, data splitting, data preprocessing, oversampling, and classification

performance [12], [31]. The Random Forest algorithm's formula is used in Eq. (8)[25], [27].

$$Gini\ (S) = 1 - \sum pi2\ k\ i = 1 \qquad (8)$$

where pi is the probability of S belonging to class i, and k is the dataset's number of classes or categories. Pi represents the proportion of data belonging to a specific class or group. The algorithm includes the following phases. [25], [31], [32]:

1. Take random samples from the database.
2. Create a decision tree for every sample. Obtain predictions from each decision tree.
3. Count the frequency of results in each class.
4. Choose the most frequent result as the final projection.

K. DEEP LEARNING (H2O)

H2O is a framework used to enable data processing and model evaluation, which has many machine learning libraries, as well as an engine for parallel processing, mathematical libraries, analysis, and deep learning with fast and scalable algorithms, while Deep Learning is an algorithm that is effective by analyzing complex problems, it is closely related to Artificial Intelligence and tries to imitate what the human brain can do, with the concept that the more the amount of data/layers, the higher the level of depth [33], [34]. The Deep Learning framework (H2O) can prevent overfitting[34].

Deep Learning(H2O) has three activation functions : Tanh, Rectified Linier , and Maxout. In this study, the function employed is Rectified Linier. Rectifier is a popular choice for activation functions in Deep Learning since it is simple, computationally efficient, and performs well in most circumstances[35]. This activation function is also ReLU (Rectified Linier Unit. Here's the formula and explanation for Rectifeir in the context of Deep Learning with H2O (Eq. (9)):

$$f(\alpha) = \max(0, \alpha) \qquad (9)$$

The activation function's output is denoted by $f(\alpha)$, while its input is α . When the input α is positif , the output is the same as the input; whwn the input is negative, the output remains zero[34]

L. EVALUATION

This research uses a confusion matrix to determine accuracy, precision, recall, and AUC (FIGURE 4). The accuracy, precision, and recall formula is used in the Eqs. (10), (11), and (12) [7], [30], [36]

$$\text{Accuracy} = \frac{(TP+TN)}{(TP + FP + FN + TN)} \qquad (10)$$

$$\text{Precision} = \frac{TP}{(TP + FP\ )} \qquad (11)$$

$$\text{Recall} = \frac{TP}{(TP + FP\ )} \qquad (12)$$

Accuracy is the number of samples that are correctly classified against the total number of samples [8]. Precision is the number of True Positive (TP) samples predicted by the total number of positive prediction samples [11](Eq. (11)). Recall is the number of correctly classified positive samples True Positive (TP) to the total number of positive samples [37](Eq. (12)).

Dataset is imbalance, and accuracy can provide false information or predictions about the dataset [38]. Accuracy and the F1 score only show the value at one threshold point that reflects the probability and does not cover the entire case. Accuracy, precision, and recall are only used as support and are not considered for the evaluation. This research will focus on the AUC value because it can describe the overall calculation of operator points [39].

Area Under the Curve (AUC) is a calculation used in various tasks, such as learning unbalanced data, learning to rank, and other things that cover the whole point [40]. The AUC interpretation criteria are > 0.5 - 0.6 (very weak), > 0.6 - 0.7 (weak), 0.7 - 0.8 (medium), 0.8 - 0.9 (good), and 0.9 - 1 (excellent)[41].

III. RESULTS

This section presents the performance of models for TikTok Shop closure sentiment analysis in Indonesia, using SVM, Decision Tree, Random Forest, and Deep Learning (H2O).

A. SUPPORT VECTOR MACHINE(SVM) PERFORMANCE

After performing preprocessing and vectorization steps, the SVM classification model was implemented based on Java with kernel dot type[24]. The accuracy, precision, recall, and AUC values are 80.37%, 96.30%, 45.61%, and 0.881 respectively, for SVM classification without SMOTE, while for SVM + SMOTE they are 77.91%, 63.64%, 85.96%, and 0.885.

B. DECISION TREE PERFORMANCE

Decision Tree uses the same Gini criteria as previous research [7]. The accuracy, precision, recall, and AUC values are 74.23%, 100%, 26.32%, and 0.646 respectively, for Decision Tree classification without SMOTE, while Decision Tree + SMOTE is 67.48%, 57.69%, 26.32%, and 0.591.

C. RANDOM FOREST PERFORMANCE

Random Forest uses the same parameters as the Decision Tree, namely the Gini criterion[7]. The accuracy, precision, recall, and AUC values are 65.03%, 0.00%, 0.00%, and 0.830 respectively for Random Forest classification without SMOTE, while for Random Forest + SMOTE, they are 79.14%, 87.10%, 47.37%, and 0.822.

D. DEEP LEARNING (H2O) PERFORMANCE

Deep Learning (H2O) is set at epoch 10 to adjust the K value in cross-validation [34]. The accuracy, precision, recall, and AUC values are 80.98%, 78.95%, 70.31%, and 0.867 respectively for Deep Learning (H2O) classification without SMOTE, while for Deep Learning (H2O) + SMOTE are 85.28%, 75.38%, 85.96%, and 0.900. From the research results, the best performance value of the model built with three classification algorithms can be determined by the AUC value. The results of the above research are presented in TABLE 7, which displays the optimal performance of the models built using SVM, Decision Tree, Random Forest, and Deep Learning (H2O) algorithms.

TABLE 7
Result in different machine learning method classification

Machine Learning Methods	Accuracy	Precision	Recall	AUC
SVM	80.37%	96.30%	45.61%	0.881
SVM + SMOTE	77.91%	63.64%	85.96%	0.885
Decision Tree	74.23%	100%	26.32%	0.646
Decision Tree + SMOTE	67.48%	57.69%	26.32%	0.591
Random Forest	65.03%	0.00%	0.00%	0.830
Random Forest + SMOTE	79.14%	87.10%	47,37%	0.822
Deep Learning (H2O)	80.98%	78.95%	70.31%	0.867
Deep Learning (H2O) + SMOTE	85.28%	75.38%	85.96%	**0.900**

FIGURE 5 shows the comparison of the best six model's performances based on AUC score. This comparison shows that the Deep Learning (H2O) + SMOTE algorithm model outperforms the other five models. In addition to SVM, Random Forest, and Deep Learning algorithms can be further developed for sentiment analysis. In contrast, the Decision Tree model did not show satisfactory performance in sentiment analysis of TikTok Shop closures in Indonesia.

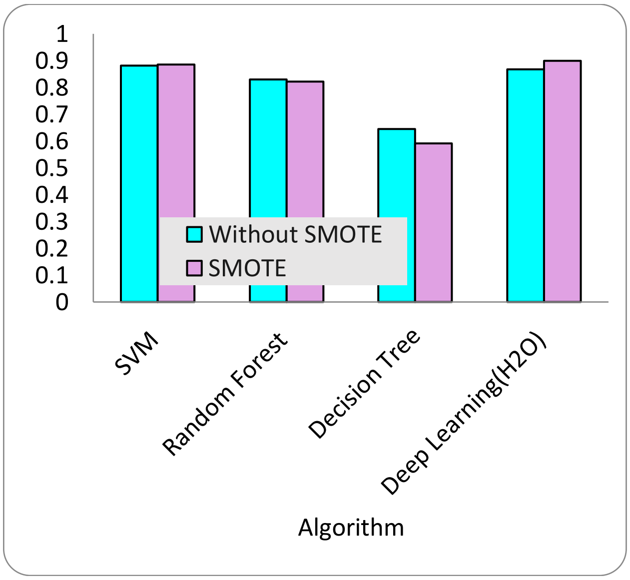

FIGURE 5. Machine learning performance comparison

IV. DISCUSSION

The performance of the sentiment analysis model built with Deep Learning (H2O) is affected by data splitting and oversampling. The best performance of this model is obtained when using the amount of data that has been added due to data oversampling.

Oversampling data does not affect the sentiment analysis model built using Decision Tree and Random Forest algorithms. TABLE 7 shows a decrease in model performance compared to without oversampling data. TABLE 8 shows a comparison of the performance of the model built in this study with the performance of models from previous studies. Previous studies used Twitter and Kaggle datasets with different amounts and topics of data. The best performance of previous studies used a model built with a Support Vector Machine (SVM) with an AUC value of 0.872 and a total amount of data of 17189 [6].

TABLE 8
Previous research on sentiment analysis

Research	Data Count	Classification	AUC
[6]	17189	**SVM**	**0.872**
		Decision Tree	0.807
[7]	1000	Random Forest	0.732
		Decision Tree	0.704
Our	1624	SVM	0.881
Research		SVM + SMOTE	0.885
		Decision Tree	0.646
		Decision Tree + SMOTE	0.591
		Random Forest	0.830
		Random Forest + SMOTE	0.822
		Deep Learning (H2O)	0.867
		Deep Learning (H2O) + SMOTE	**0.900**

A comparison between the two previous studies shows that the method proposed in this study has the potential to outperform or achieve similar results to previous studies. SVM and Deep Learning (H2O) achieve greater AUC values in this study than in the previous one [6] with more data than in [7], SVM, Deep Learning (H2O), and Random Forest get greater AUC values in this study. As a result, using 1624 data, which is significantly less than the 17189 data in the research [6]and more than 1000 data in the study [7], SVM, Deep Learning (H2O), and Random Forest with and without SMOTE perform better.

This research introduces the Deep Learning (H2O) algorithm as a new aspect of sentiment analysis research and incorporates the use of the Python programming language for crawling data, manually labeling, and using RapidMiner tools to facilitate the preprocessing process until the evaluation stage.

However, this research has limitations and shortcomings. Oversampling of data carried out only increases the AUC value in the Deep Learning (H2O) and SVM algorithms, but instead decreases the AUC value in the Decision Tree and Random Forest algorithms. The less-than-optimal performance of sentiment analysis of TikTok Shop closures in Indonesia is due to the problem of unbalanced data and the relatively small amount of data. Table IX also indicates poor performance of the Decision Tree with 1624 data in this analysis, which is significantly less than 17189 data in the study[6]and more than 1000 data in the study[7].

The sentiment analysis model developed in this research will have significant implications for social e-commerce due to its potential utilization by social media analysts. Applying this method allows social media analysts to determine user comfort with a platform, by reviewing the community's response to the platform and the cases that are being faced by the platform.

V. CONCLUSION

Data obtained from Twitter regarding the closure of TikTok Shop in Indonesia is considered unstructured and requires weighting (TF-IDF) to produce structured data suitable for machine learning algorithms. It was also discovered that the dataset was not balanced therefore, oversampling with SMOTE was required to address these problems, and dividing data was utilized to divide training data and testing data with a ratio of (90:10). With the number of 1624, which is less than 17189 and more than 1000, the best algorithm performance is SVM, Random Forest and Deep Learning (H2O). Unlike the Decision Tree, the more datasets, the more performance will also increase.

To determine if SMOTE is an efficient method for dealing with data imbalances, prediction models with and without SMOTE are treated. Thus, this research includes six prediction model developments that use three different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Random Forest, and Deep Learning (H2O). SMOTE was able to improve the performance of Support Vector Machine and Deep Learning (H2O) algorithms, but not Decision Tree and Random Forest algorithms. Support Vector Machine's AUC of 0.881 increased to 0.885. Deep Learning (H2O) increased from 0.867 to 0.900. Decision Tree decreased from 0.646 to 0.591, and Random Forest also decreased from 0.830 to 0.822. It can be concluded that the best model for this research is Deep Learning (H2O) + SMOTE with an AUC value that can enter the excellent category.

This research still has some limitations when viewed from perspective of the performance of the model, which produces an AUC value below 0,7. This sub-optimal model performance can be caused by the model being built using a small amount of data.

Given these limitations and shortcomings, it is recommended that further research be conducted to gather fresh data that includes minority class data, ensuring that the dataset is balanced and has more information. Furthermore, future researchers should investigate the use of classification approaches with other algorithms or suggest ways to improve the present algorithms in this study's sentiment analysis index.

ACKNOWLEDGMENT

Researchers would like to thank all the individuals from the Computer Science Department, Lambung Mangkurat University, who have contributed to the completion of this research project. Their input and suggestions provided by several parties have significantly improved the quality of the research. Researchers are also indebted to our project team members for their collaboration and hard work, which have contributed significantly to the research's success.

REFERENCES

[1] S. Hu, U. Akram, F. Ji, Y. Zhao, and J. Song, “Does social media usage contribute to cross-border social commerce? An empirical

evidence from SEM and fsQCA analysis," *Acta Psychol (Amst)*, vol. 241, Nov. 2023, doi: 10.1016/j.actpsy.2023.104083.

[2] M. Alviazra Virgananda, I. Budi, and R. Randy Suryono, "Purchase Intention and Sentiment Analysis on Twitter Related to Social Commerce." [Online]. Available: www.ijacsa.thesai.org

[3] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1. Springer, Dec. 01, 2021. doi: 10.1007/s13278-021-00776-6.

[4] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, Feb. 2018, doi: 10.3390/mca23010011.

[5] O. Czeranowska *et al.*, "Migrants vs. stayers in the pandemic – A sentiment analysis of Twitter content," *Telematics and Informatics Reports*, vol. 10, Jun. 2023, doi: 10.1016/j.teler.2023.100059.

[6] H. Cam, A. V. Cam, U. Demirel, and S. Ahmed, "Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers," *Heliyon*, vol. 10, no. 1, p. e23784, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23784.

[7] V. Umarani, A. Julian, and J. Deepa, "Sentiment Analysis using various Machine Learning and Deep Learning Techniques," *Journal of the Nigerian Society of Physical Sciences*, vol. 3, no. 4, pp. 385–394, Nov. 2021, doi: 10.46481/jnsps.2021.308.

[8] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 662–668. doi: 10.1016/j.procs.2016.05.251.

[9] A. T. Gurmu, A. Krezel, and M. N. Mahmood, "Analysis of the causes of defects in ground floor systems of residential buildings," *International Journal of Construction Management*, vol. 23, no. 2, pp. 268–275, 2023, doi: 10.1080/15623599.2020.1860636.

[10] M. Z. Naser, "Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering problems," *Journal of Infrastructure Intelligence and Resilience*, vol. 2, no. 1, Mar. 2023, doi: 10.1016/j.iintel.2023.100028.

[11] V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in *Proceedings of the First International Conference on Information Technology and Knowledge Management*, PTI, Jan. 2018, pp. 71–75. doi: 10.15439/2017km46.

[12] K. Purwandari, R. B. Perdana, J. W. C. Sigalingging, R. Rahutomo, and B. Pardamean, "Automatic Smart Crawling on Twitter for Weather Information in Indonesia," *Procedia Comput Sci*, vol. 227, pp. 795–804, 2023, doi: 10.1016/j.procs.2023.10.585.

[13] S. Leorna and T. Brinkman, "Human vs. machine: Detecting wildlife in camera trap images," *Ecol Inform*, vol. 72, Dec. 2022, doi: 10.1016/j.ecoinf.2022.101876.

[14] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," *IOP Conf Ser Mater Sci Eng*, vol. 1090, no. 1, p. 012053, Mar. 2021, doi: 10.1088/1757-899x/1090/1/012053.

[15] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf Syst*, vol. 121, Mar. 2024, doi: 10.1016/j.is.2023.102342.

[16] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/4832864.

[17] B. T. Pham *et al.*, "A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil," *Sustainability (Switzerland)*, vol. 12, no. 6, pp. 1–16, Mar. 2020, doi: 10.3390/su12062218.

[18] X. Xiao *et al.*, "Treatment initiation prediction by EHR mapped PPD tensor based convolutional neural networks boosting algorithm," *J Biomed Inform*, vol. 120, Aug. 2021, doi: 10.1016/j.jbi.2021.103840.

[19] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, doi: 10.1016/j.jksuci.2021.01.014.

[20] K. Suryadi, R. Herteno, S. Wahyu Saputro, M. Reza Faisal, R. Adi Nugroho, and M. Kevin Suryadi, "ShareAlike 4.0 International License (CC BY-SA 4.0). How to cite: Mulia Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect A Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 137–147, 2024, doi: 10.35882/jeeemi.v6i2.375.

[21] C. Zhang, J. Song, Z. Pei, and J. Jiang, "An Imbalanced Data Classification Algorithm of De-noising Auto-Encoder Neural Network Based on SMOTE", doi: 10.1051/conf/2016.

[22] G. Baron and U. Stanczyk, "Standard vs. non-standard cross-validation: Evaluation of performance in a space with structured distribution of datapoints," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 1245–1254. doi: 10.1016/j.procs.2021.08.128.

[23] A. M. Peco Chacón, I. Segovia Ramírez, and F. P. García Márquez, "K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection," *Sustainable Futures*, vol. 6, Dec. 2023, doi: 10.1016/j.sftr.2023.100132.

[24] S. Ma, W. Cao, S. Jiang, J. Hu, X. Lei, and X. Xiong, "Design and implementation of SVM OTPC searching based on Shared Dot Product Matrix," *Integration*, vol. 71, pp. 30–37, Mar. 2020, doi: 10.1016/j.vlsi.2019.11.007.

[25] P. A. Riadi, M. R. Faisal, D. Kartini, R. A. Nugroho, D. T. Nugrahadi, and D. B. Magfira, "A Comparative Study of Machine Learning Methods for Baby Cry Detection Using MFCC Features," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, Jan. 2024, doi: 10.35882/jeeemi.v6i1.350.

[26] M. I. Mazdadi, I. Budiman, and R. Herteno, "IMPLEMENTATION OF INFORMATION GAIN AND PARTICLE SWARM OPTIMIZATION ON SENTIMENT ANALYSIS OF COVID-19 HANDLING USING K-NN," *Jurnal Informatika dan Komputer) Accredited KEMENDIKBUD RISTEK*, vol. 6, no. 1, 2023, doi: 10.33387/jiko.v6i1.5260.

[27] M. T. Hidayat, M. R. Faisal, D. Kartini, F. Indriani, I. Budiman, and T. H. Saragih, "Comparison of Machine Learning Performance on Classification of COVID-19 Cough Sounds Using MFCC Features Porównanie wydajności uczenia maszynowego w zakresie klasyfikacji odgłosów kaszlu COVID-19 przy użyciu funkcji MFCC," 2023.

[28] S. Lee, C. Lee, K. G. Mun, and D. Kim, "Decision Tree Algorithm Considering Distances between Classes," *IEEE Access*, vol. 10, pp. 69750–69756, 2022, doi: 10.1109/ACCESS.2022.3187172.

[29] E. Dritsas and M. Trigka, "Machine Learning Techniques for Chronic Kidney Disease Risk Prediction," *Big Data and Cognitive Computing*, vol. 6, no. 3, Sep. 2022, doi: 10.3390/bdcc6030098.

[30] N. H. Arif *et al.*, "Approach to ECG-based Gender Recognition Using Random Forest Algorithm," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 107–115, 2024, doi: 10.35882/jeeemi.v6i2.363.

[31] D. Si, W. Hu, Z. Deng, and Y. Xu, "Fair hierarchical clustering of substations based on Gini coefficient," *Global Energy Interconnection*, vol. 4, no. 6, pp. 576–586, Dec. 2021, doi: 10.1016/j.gloei.2022.01.009.

[32] P. Kulkarni, S. Umarani, V. Diwan, V. Korde, and P. P. Rege, "Child Cry Classification - An Analysis of Features and Models," in *2021 6th International Conference for Convergence in Technology, I2CT 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021. doi: 10.1109/I2CT51068.2021.9418129.

[33] D. Suleiman and G. Al-Naymat, "SMS Spam Detection using H2O Framework," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 154–161. doi: 10.1016/j.procs.2017.08.335.

[34] A. Candel, E. Ledell, and A. Bartz, "Deep Learning with H2O." [Online]. Available: http://h2o.ai/resources/

[35] A. Maniatopoulos and N. Mitianoudis, "Learnable Leaky ReLU (LeLeLU): An Alternative Accuracy-Optimized Activation Function," *Information (Switzerland)*, vol. 12, no. 12, Dec. 2021, doi: 10.3390/info12120513.

[36] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[37] D. López, I. Aguilera-Martos, M. García-Barzana, F. Herrera, D. García-Gil, and J. Luengo, "Fusing anomaly detection with false positive mitigation methodology for predictive maintenance under multivariate time series," *Information Fusion*, vol. 100, Dec. 2023, doi: 10.1016/j.inffus.2023.101957.

[38] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," in *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, Elsevier, 2020, pp. 83–106. doi: 10.1016/B978-0-12-818366-3.00005-8.

[39] A. M. Carrington *et al.*, "Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation," Mar. 2021, doi: 10.1109/TPAMI.2022.3145392.

[40] W. Gao, L. Wang, R. Jin, S. Zhu, and Z. H. Zhou, "One-pass AUC optimization," *Artif Intell*, vol. 236, pp. 1–29, Jul. 2016, doi: 10.1016/j.artint.2016.03.003.

[41] A. Br Haloho, Z. Hafy, and A. Annisa Rizki, "ISSN 2598 0580 Bioscientia Medicina Sensitivity and Specificity of Urine N-Acetyl-β-D-Glucosaminidase as an Early Biomarker For Acute Kidney Injury." [Online]. Available: www.bioscmed.com

AUTHORS BIOGRAPHY

Noor Zalekha Al Habesyah is from Banjarbaru, South Kalimantan. Since 2020, she has been studying Computer Science at Lambung Mangkurat University. Her current research focus is on software engineering and text mining. This program provides opportunity for her to pursue her interests in this profession. She chose this particular interest since she was fascinated by software engineering and had a strong interest in the field. In addition, her thesis requires her to undertake research on sentiment analysis of software or features. This study seeks to determine the performance of machine learning classification for sentiment analysis of TikTok business closures in Indonesia.
Email address: 2011016320013@mhs.ulm.ac.id.

Rudy Herteno is currently a lecturer in the Faculty of Mathematics and Natural Science, at Lambung Mangkurat University. He received his bachelor's degree in Computer Science from Lambung Mangkurat University and a master's degree in Informatics from STMIK Amikom University. His research interests include software engineering, software defect prediction, and deep learning.
Email address: rudy.herteno@ulm.ac.id

Fatma Indriani is a lecturer in the Department of Computer Science at Lambung Mangkurat University. Her research interests include data science and artificial intelligence. She earned her Bachelor's Degree in Informatics Engineering from Institut Teknologi Bandung. Subsequently, she completed her Master's studies at Monash University, Australia, and her PhD studies at Kanazawa University, Japan. Email address : f.indriani@ulm.ac.id

Irwan Budiman was born in Banjarmasin. He is a lecturer at Lambung Mangkurat University and currently serves as the Coordinator in the Department of Computer Science, Faculty of Mathematics and Natural Sciences**.** He earned his Bachelor's Degree in Informatics Engineering from Islam Indonesia University, Yogyakarta. Subsequently, he completed his Master's studies in information systems at Diponegoro University, Semarang. His research interests include data mining, human-computer interaction, applied business intelligence, and e-government. Email address: irwan.budiman@ulm.ac.id

Dwi Kartini received her Bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia "YPTK" Padang, Indonesia. She is a lecturer too in the Department of Computer Science. She instructs various subjects such as linear algebra, discrete mathematics, research methods and others Her research interests include the applications of Artificial Intelligence and Data Mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia.