Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 17(1) 2024: 25-35
http://ejournal.undip.ac.id/index.php/media_statistika

MULTICLASS CLASSIFICATION OF MARKETPLACE PRODUCTS WITH MACHINE LEARNING

Farhan Satria Aditama[1,3], Dewi Krismawati[2], Setia Pramana[3]
[1] Directorate of Statistical Dissemination, BPS Statistics Indonesia, Jakarta, Indonesia
[2] Directorate of Analysis and Statistics Development, BPS Statistics Indonesia, Jakarta, Indonesia
[3] Politeknik Statistika STIS, Jakarta, Indonesia

e-mail: farhan.satria@bps.go.id

DOI: 10.14710/medstat.17.1.25-35

Article Info:
Received: 8 January 2024
Accepted: 20 September 2024
Available Online: 14 October 2024

Keywords:
Machine Learning, Marketplace, Multiclass Classification.

Abstract: The use of marketplace data and machine learning in the collection of commodity data can provide an opportunity for Statistics Indonesia to complete the commodity directories for various surveys. This research adopts machine learning to train a product classification model based on existing datasets to predict whether a new dataset falls into which KBKI category. The dataset contains more than 32,000 products from 26 classes consisting of product data from two biggest marketplaces in Indonesia. Algorithms used for classification include Random Forests (RF), Support Vector Machines (SVM), and Multinomial Naive Bayes (MNB). Results indicate that MNB is the most effective algorithm when considering the trade-off between accuracy and processing time. MNB achieved the highest micro-average F1 scores, with 91.8% for Tokopedia and 95.4% for Shopee, and has the fastest execution time approximately 5 seconds.

1. INTRODUCTION

BPS Statistics Indonesia has initiated the utilization of big data sources to modernize its statistics business process (Saleh et al., 2019). Big data offers an innovative source of information that provides deeper insight into the production of official statistics (Badan Pusat Statistik, 2020). Commodity data collection by BPS (Badan Pusat Statistik) is very important to obtain accurate and comprehensive information regarding the production and consumption of certain commodities in Indonesia. Collecting data from marketplaces offers a more efficient and reliable approach, especially in situations requiring a quick response. Marketplace data also provides valuable insights into ongoing phenomena (Srimulyani et al., 2021).

However, utilizing marketplace data for official statistics presents several challenges, particularly in the classification of products into standard categories such as the Standard Classification of Indonesian Commodities (KBKI). The KBKI system is used to categorize various goods and services traded in Indonesia, facilitating the organization of trade data and supporting the collection of statistical information (Badan Pusat Statistik, 2012). Effective

classification is critical for ensuring the accuracy and relevance of the data used in economic analysis.

The primary challenges in product classification include the unbalanced distribution of product categories, inconsistent product descriptions provided by sellers, and the high dimensionality of the classification task due to the large number of categories. These challenges complicate the classification process, increasing the need for sophisticated methods to handle complex and large-scale datasets (Yu et al., 2018).

To address these challenges, this research proposes the development of machine learning models tailored to the characteristics of marketplace data. By applying algorithms such as Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Random Forest (RF), this research aims to improve the accuracy and efficiency of product classification based on the KBKI system. These algorithms are selected for their proven effectiveness in text classification tasks, with each offering distinct advantages in handling different aspects of the classification problem (Laksana & Purwarianti, 2014).

This study contributes to the field of official statistics by exploring the application of machine learning techniques to enhance the processing of big data for commodity classification. It represents a pioneering effort in utilizing marketplace data for the classification of products according to the KBKI system, providing a foundation for future research and potential improvements in the statistical processes of BPS Statistics Indonesia

There are several studies related to the use of multiclass classification, such as (Arusada et al., 2017), who uses Naive Bayes and Support Vector Machine as classifiers. They showed that the training data optimization strategy can provide good performance for multiclass text classification models. In addition, Ibnu & Rachmatullah (2022) showed that the best performance in multiclass classification uses the random forest classifier method. On the other hand, Multinomial Naive Bayes (MNB) is efficient and quick to build, making it suitable for situations where speed is crucial (Laksana & Purwarianti, 2014).

2. LITERATURE REVIEW

2.1. Multiclass Classification Model

Model building is performed using the Scikit-Learn library in Python. The classification model used uses supervised learning methods, including the SVM, RF, and MNB algorithms. Below is an explanation of each model:

1. Support Vector Machine (SVM)

   SVM is one of the most popular supervised learning algorithms, commonly used for classification problems (Awad & Rahul Khanna, 2015). SVM operates by finding a hyperplane that maximizes the margin between classes. The data points closest to the hyperplane are referred to as support vectors (Korde, 2012). An illustration of classification using SVM is shown in Figure 1.

   Data classification is not always straightforward in a linear model. SVM can utilize non-linear kernels to maximize the hyperplane margin. SVM was chosen for its strong generalization, broad applicability, and effective handling of high-dimensional data, making it well-suited for complex datasets. Table 1 provide popular and frequently used kernel functions.

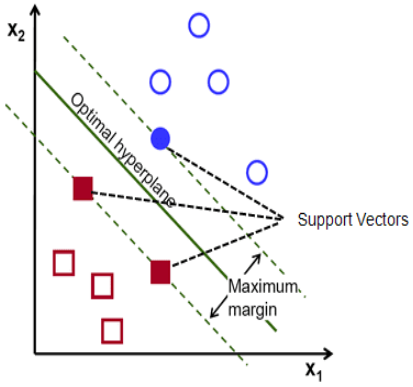

Figure 1. Illustration of classification using SVM
Source: (UNECE, 2022)

Table 1. Kernel Functions Commonly Used

Kernel Function	Function Formula
Linear	$K(x_i, x_j) = \langle x_i, x_j \rangle$
Polynomial	$K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d$
Gaussian Radial Basis Function	$exp\ exp\left(-\gamma \|x_i - x_j\|^2\right)$ with $\gamma > 0$ or $\gamma = \frac{1}{2\sigma^2}$

2. Random Forest (RF)
   RF algorithm is a variants of bagging ensemble method (Breiman, 2001). RF is a combination of decision trees, where each tree depends on random vectors sampled independently with the same distribution. Each tree generates its own prediction, and the final class prediction is determined by majority voting among all the trees' predictions.
3. Multinomial Naïve Bayes (MNB)
   MNB is a classification method in machine learning that uses probability theory and is based on Bayes' theorem, which was introduced by Thomas Bayes (Suyanto, 2019). This method considers the word count in a document, assuming that the occurrence of words is independent, without regard to word order or context. Based on this assumption, Multinomial Naive Bayes estimates the class probability of a document by analyzing word occurrences. In classification, it predicts the document's class by selecting the one with the highest estimated probability using the MNB model.

2.2. Text Preprocessing

Preprocessing is a crucial step in data analysis, used to clean and prepare data before further analysis. Raw data often has irregular structures, contains noise, or is in a format unsuitable for analysis. Preprocessing is performed to remove noise from the data, thereby enhancing classification performance (Işik & Dağ, 2020). The processes involved in this stage include:

1. Case Folding: all uppercase letters in the dataset are converted to lowercase using the lower string module. This step ensures consistency in the case of the letters, which allows the model to recognize similar words as being identical and to avoid redundancy.
2. Data Cleaning: attributes and elements that do not support the commodity classification process are removed. This includes removing numbers, punctuation marks, and white spaces. Removing the numbers reduces the dimensionality of the dataset, making subsequent classification easier.
3. Tokenization: Tokenization involves breaking down sentences in the dataset into individual words. This process is facilitated by the *"re"* library, which allows the model to analyze each word separately.

4. Stopword Removal: words that appear frequently in the dataset but have minimal impact on the performance of the classification model are removed and replaced with spaces.

2.3. Feature Extraction

Feature extraction converts data from string type to numeric type to enable the modeling process. Datasets that have gone through a preprocessing process become input material at this stage (Xu & Wu, 2014). For feature extraction in this research, the Term Frequency-Inverse Document Frequency (TF-IDF) technique from the *Scikit-learn* library was used.

TF-IDF has the advantage of representing word weights by assigning higher weights to words that are rare across the document collection (Ansari et al., 2020;). The TF-IDF formula for a term i in document j is expressed as follows:

$$w_{i,j} = tf \times log(\frac{N}{df_i}) \quad (1)$$

where $tf_{i,j}$= number of occurences of i in j, df_i = number of documents containing i, and N = total number of documents.

2.4. SMOTE

SMOTE is an oversampling technique designed to balance data by generating synthetic data from minority classes. By generating additional synthetic data points, SMOTE helps overcome the problem of overfitting and improves the performance of classification models by addressing class imbalance in the dataset. Several studies have shown that the SMOTE method can improve the accuracy of classifications as mentioned in studies (Chau, V.T., & Phung, 2013; Ünal et al., 2019). According to (Chawla et al., 2002), the SMOTE approach generates fictional data based on the k-nearest neighbor (KNN) and multiplies the amount of data in the minor class to be equivalent to the major class.

2.5. Cross Validation

Cross-validation is a widely used evaluation method in statistical modelling and machine learning. It involves dividing the data into multiple distinct subsets, or folds. In each iteration, one subset is designated as the testing set, while the remaining subsets are used as the training set. This process is repeated for each fold, allowing the model to be evaluated across different portions of the data, ensuring a more robust assessment.

Grid cross-validation is a specific approach within cross-validation that focuses on optimizing hyperparameters. It involves creating a grid of all possible combinations of hyperparameter values. For each combination, the model is evaluated using cross-validation. This systematic search helps identify the best set of hyperparameters for the model.

2.6. Evaluation Model

Table 2. Confusion Matrix Illustration

Confusion Matrix		Predicted		
		Commodity 1	Commodity 2	Commodity 3
Actual	Commodity 1	N_{11}	N_{12}	N_{13}
	Commodity 2	N_{21}	N_{22}	N_{23}
	Commodity 3	N_{31}	N_{32}	N_{33}

Source: (Febriantono et al., 2020; Idris, 2018)

Model evaluation is performed using the confusion matrix, focusing on metrics such as accuracy, precision, recall, and F1-score (Table 2). Cross-validation is used to assess the performance of the model. Specifically, a 10-fold cross-validation is performed using the

Grid Search Cross Validation method. In addition, an iteration process is applied during the training and testing of the classification model to ensure more stable and representative results.

Based on the confusion matrix presented, we can derive performance metrics to evaluate the effectiveness of the classification model using the test data. The accuracy rate is the ratio of correct predictions to the number of products

$$Accuracy = \frac{Total\ product\ correct\ predictions}{Total\ number\ of\ dataset} \quad (2)$$

Precision, also known as positive predictive value, represents the proportion of positive correct predictions to the total number of positive predictions.

$$Precision = \frac{products\ in\ commodities\ n\ that\ are\ correctly\ prediction}{predicted\ total\ product\ on\ commodity\ n} \quad (3)$$

Recall or sensitivity, also known as the true positive rate, represents the ratio of positive correct predictions to the total number of actual positive instances in the dataset.

$$Recall = \frac{products\ in\ commodities\ n\ that\ are\ correctly\ prediction}{actual\ total\ product\ on\ commodity\ n} \quad (4)$$

The F1 score is the harmonic mean of recall and precision.

$$F1\ Score = \frac{2\times(recall\times precision)}{recall+precision} \quad (5)$$

In the case of a data set that is not evenly distributed, a better approach is to combine the metrics into a composite metric such as F1 score (UNECE, 2022). Micro-averaging is then used to calculate performance metrics such as precision, recall, and F1 score simultaneously in multi-class classification.

3. MATERIAL AND METHOD

3.1. Dataset

There are two datasets used in this paper, i.e., the product Tokopedia dataset and Shopee dataset. The dataset collected is primary data with random sampling. For the Tokopedia data used in this study consists of product information collected in July 2022, with a total of 4.3 million product samples. The Shopee data was obtained from a previous study conducted by (Ghozy & Pramana, 2020), and the product data used was collected in April 2020. The successful collection of Shopee data resulted in 228 thousand product samples. Although the data from Tokopedia and Shopee were collected at different points in time, this time difference does not affect the study as the product name attribute is used for classification purposes.

3.2. Data Labeling

In this research, the data consisted of 26 labels, namely (1) Products from agriculture, horticulture and plantations, (2) Live Animals and animal products (excluding meat), (4) Fish and other fishery products, (16) Other minerals, (21) Meat, fish, fruit, vegetables, oils and fats, (22) Dairy products and egg products, (23) Milled grain products, starch and starch products, other food products, (24) Drink, (26) Weaving/knitting yarn and twine; woven fabrics and tufted textile fabrics, (27) Textile goods other than clothing, (28) Knitted or crocheted fabric; clothes, (29) leather and leather products; footwear, (31) Products from wood, cork, straw and woven materials, (32) Pulp, paper and paper products; printed matter and related items, (35) other chemical products; artificial fiber, (36) rubber and plastic products, (37) Glass and glass products and other non-metallic products, etc., (38) Household furniture; other items listed that can be moved, (42) Fabricated Metal Products, except

machinery and equipment, (43) General purpose machine, (44) Machines for special purposes, (45) Office, accounting and computing machines, (46) Electrical machines and devices, (47) Radio, television and communication equipment and their accessories, (48) Equipment for medical, precision and optical applications, watches and clocks, (49) Transportation equipment.

3.3. Method

The research aims to examine the construction of a text classification model to classify marketplace products according to the *KBKI*, and the overall workflow for this proposed approach is shown in Figure 2.

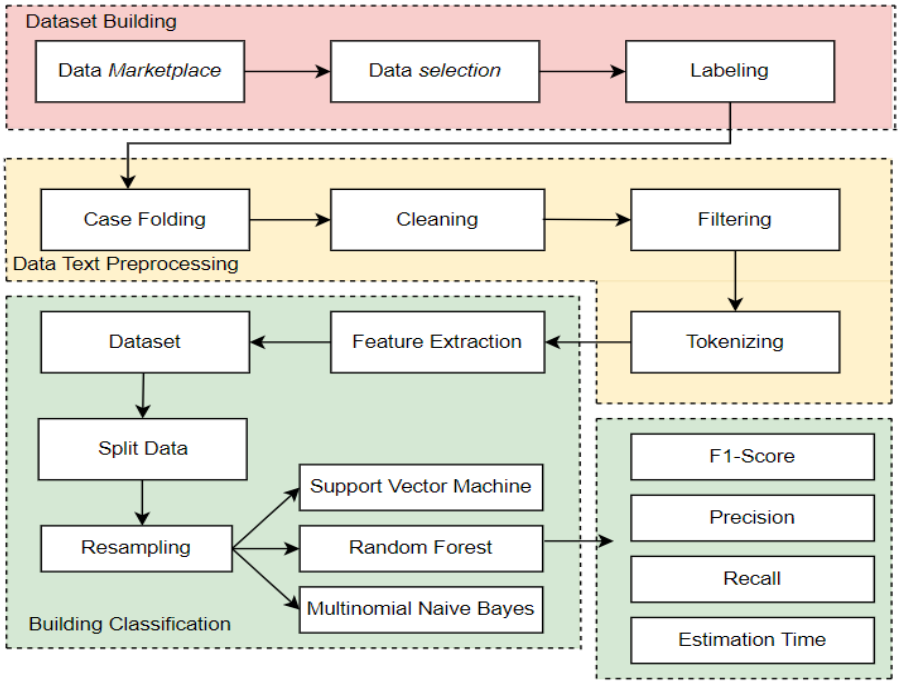

Figure 2. Workflow for Text Classification for Classifying Marketplace Product

The algorithm for modeling can be describe as follows:
Dataset Building: Begin by manually labeling product titles according to the two-digit KBKI code. The labeling process is guided by KBKI keywords to ensure consistency.
Data Text Preprocessing: Text preprocessing is a fundamental step that transforms raw text into a clean and structured format suitable for machine learning models. This process begins with case folding, data cleaning, tokenization and stopword removal.
Building Classification: (1) Convert the cleaned and tokenized text into a matrix of features using Term Frequency-Inverse Document Frequency (TF-IDF); (2) Randomly split the dataset into training, validation, and test sets with a ratio of 64:16:20, respectively. To enhance the model's learning and generalization capabilities, combine training and validation data from multiple platforms, such as Tokopedia and Shopee; (3) Apply the Synthetic Minority Oversampling Technique (SMOTE) to the training data to address class imbalance and ensure that minority classes are adequately represented; (4) Utilize machine learning classification models, including Support Vector Machine (SVM), Random Forest (RF), and Multinomial Naive Bayes (MNB); (5) Perform grid search cross-validation to optimize model parameters and improve performance; (6) Employ 10-fold cross-validation to prevent overfitting and ensure that the model results are reliable and robust; (7) Evaluate the model

using a confusion matrix to compute various performance metrics, including accuracy, precision, recall, F1 score, and execution time.

4. RESULTS AND DISCUSSION

The data collection is done by exporting product data from Tokopedia and Shopee. The dataset collected consists of 4,356,226 product samples from Tokopedia and 228,726 product samples from Shopee. The next stage is manual labeling based on the KBKI classification guidelines. This process categorizes product names based on the two-digit KBKI code. researchers have completed labeling tasks for a total of 32,932 products.

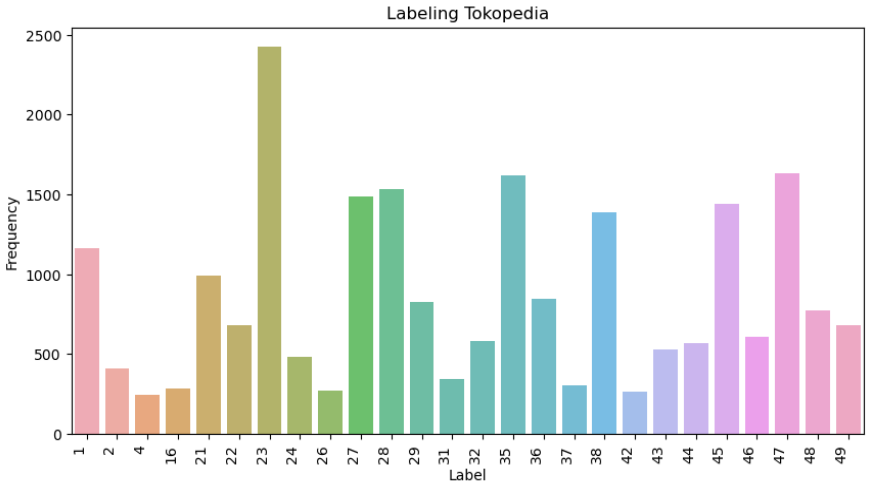

Figure 3. Tokopedia Product Labeling Results

Figure 2 shows the number of Tokopedia product samples that have been labeled. Of the total 22,318 products that have been labeled, there is an imbalance in the distribution of data in each class. In class 23 (Grain products, starch and starch products, other food products) there are 2423, while in class 4 (Fish and other fishery products) there are 245.

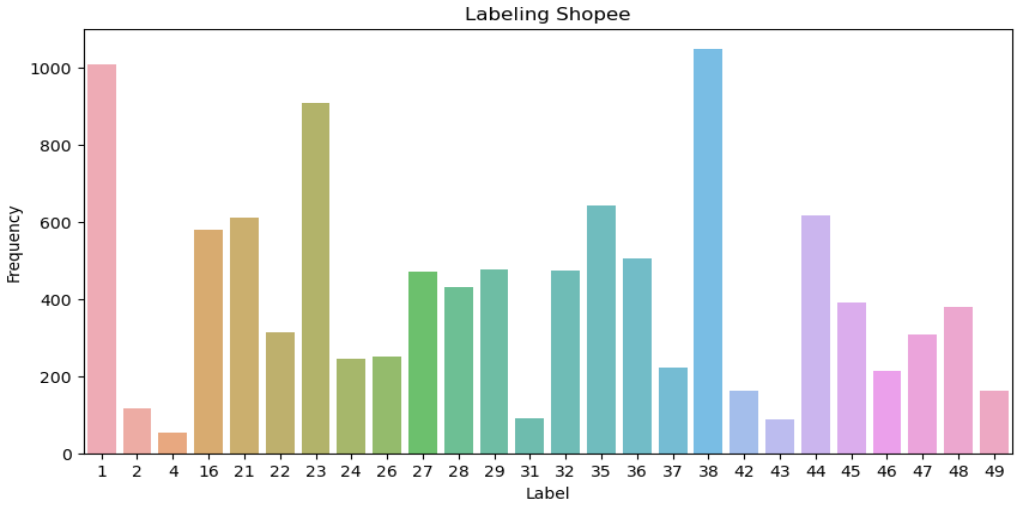

Figure 4. Shopee Product Labeling Result

Figure 3 shows the number of Shopee sample products that have been labeled. Of the total 10,772 products that have been labeled, there is an imbalance in the distribution of data in each class. In class 38 (Household furniture; other movable goods), there are 1048, while class 4 (Fish and other fishery products) has 54.

Preprocessing of marketplace product data text carried out in this research is case folding, cleaning, tokenizing, and stop word removal. An example of the results of the preprocessing stage can be seen in Table 3.

Table 3. Text Preprocessing of Product Marketplace

Product Name	Case folding	Cleaning	Tokenizing	Stopword removal
Timbangan badan digital eb 9362 onemed	timbangan badan digital eb 9362 onemed	timbangan badan digital eb onemed	['timbangan', 'badan', 'digital', 'eb', 'onemed']	['timbangan', 'badan', 'digital', 'eb', 'onemed']
Keripik pisang suseno ambon 500gram kripik lampung	keripik pisang suseno ambon 500gram kripik lampung	keripik pisang suseno ambon gram kripik lampung	['keripik', 'pisang', 'suseno', 'ambon', 'gram', 'kripik', 'lampung']	['keripik', 'pisang', 'suseno', 'ambon', 'kripik', 'lampung']
Kursi bakso plastik kursi makan model rotan olymplast	kursi bakso plastik kursi makan model rotan olymplast	kursi bakso plastik kursi makan model rotan olymplast	['kursi', 'bakso', 'plastik', 'kursi', 'makan', 'model', 'rotan', 'olymplast']	['kursi', 'bakso', 'plastik', 'kursi', 'makan', 'model', 'rotan', 'olymplast']

Model evaluation results for the SVM, RF and MNB algorithms are presented in Table 4.

Table 4. The Result of The Multiclass Using 10- Fold CV

Model	Precision	Recall	F1-Score
SVM Tokopedia	94.9%	95.0%	94.9%
SVM Shopee	96.9%	97.0%	96.9%
RF Tokopedia	71.5%	78.6%	72.2%
RF Shopee	78.7%	83.5%	79.4%
MNB Tokopedia	91.8%	91.9%	91.8%
MNB Shopee	95.4%	95.5%	95.4%

Table 5. The Result of The Multiclass Classification Results Based on Execution Time

Model	Execution Time (time)
Support Vector Machine (SVM)	6 hours 46 minutes
Random Forest (RF)	2 hours 35 minutes
Multinomial Naïve Bayes (MNB)	5 seconds

Table 4 shows the results of determining the best model using a 10-fold CV. Test data shows that the model using the SVM algorithm has the best performance. This conclusion is based on the average micro F1 SVM score, which achieved the highest scores for Tokopedia and Shopee test data, at 94.9% and 96.9%, respectively. Table 5 presents the execution time for each model, MNB has the fastest execution time, taking only about 5 seconds to build and predict the test data set.

Table 6. Examples of Incorrect Classification Results for Marketplace Products

No	Dataset	Product Name	Label	Prediction
1	Shopee	buah pisang tanduk sukabumi	1	23
2	Shopee	toples salak dari kayu jati asli	38	23
3	Tokopedia	buah cempedak matang manis	1	23

After applying the classification model, it is important to check and compare the actual labels with the predicted labels. Example in Table 6 shows the results of a product name with an error in the model prediction. Table 6 shows examples of misclassification of

marketplace product names. The product name *"buah pisang tanduk sukabumi"* should be labelled 1 (Products from agriculture, horticulture and plantations). However, after applying the model, the banana is predicted to fall under label 23 (Milled grains, starch and starch products, other food products). Further investigation of 23's labels revealed products named 'banana chips' and 'fried banana flour', both of which are banana derivative products. Evaluating data errors can help improve modelling results by increasing the data set for misclassified classes.

The SVM and MNB models have better performance compared to the RF model. SVM and MNB have their respective advantages; In this research, MNB had a faster training time than SVM. On the other hand, SVM algorithm has the best performance. If the dataset later becomes larger, researchers recommend using MNB because it has a shorter training time. This is in line with (Baeza-Yates & Liaghat, 2017) MNB showing that the best algorithms achieve a balance between quality and efficiency.

Identifying patterns of misclassification may indicate the need for improved feature engineering or the inclusion of additional features. Frequent misclassifications between categories might also suggest a need for clearer labeling criteria or enhanced keyword mapping to better differentiate between categories.

5. CONCLUSION

Based on the trade-off between accuracy and processing time, the Multinomial Naive Bayes (MNB) model was identified as the most effective model in this study. The MNB model achieved the highest micro-average F1 scores, reaching 91.8% on Tokopedia test data and 95.4% on Shopee test data. Additionally, it demonstrated superior efficiency by completing the model building and prediction process in approximately 5 seconds.

ACKNOWLEDGMENT

We would like to thank the Politeknik Statistika STIS and the BPS Statistics Indonesia for providing access to data.

REFERENCES

Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., & Singh, K. P. (2020). Analysis of Political Sentiment Orientations on Twitter. *Procedia Computer Science*, *167*, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

Arusada, M. D. N., Putri, N. A. S., & Alamsyah, A. (2017). Training Data Optimization Strategy for Multiclass Text Classification. *2017 5th International Conference on Information and Communication Technology, ICoIC7 2017, February*. https://doi.org/10.1109/ICoICT.2017.8074652

Awad, M., & Rahul Khanna. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. USA: Springer Nature.

Badan Pusat Statistik. (2012). *Klasifikasi Baku Komoditas Indonesia (KBKI) 2012* (Vol. 1). https://www.bps.go.id/publication/2012/11/30/7c6cba9683c26ffbcdf5561e/klasifikasi-baku-komoditas-indonesia--kbki--2012-buku-1.html

Badan Pusat Statistik. (2020). *Kajian Big Data Sebagai Pelengkap Data Dan Informasi Statistik Ekonomi*.

Baeza-Yates, R., & Liaghat, Z. (2018). Quality-Efficiency Trade-Offs in Machine Learning For Text Processing. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 897–904. https://doi.org/10.1109/BigData.2017.8258006

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chau, V.T., & Phung, N. H. (2013). Imbalanced Educational Data Classification: An Effective Approach with Resampling and Random Forest. *The 2013 RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, 135-140. https://doi.org/10.1109/RIVF.2013.6719882

Chawla, N. V, K. W. Bowyer, L. O. H., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(1), 321-357. https://doi.org/10.1613/jair.953

Febriantono, M. A., Pramono, S. H., Rahmadwati, & Naghdy, G. (2020). Classification Of Multiclass Imbalanced Data Using Cost-Sensitive Decision Tree c5.0. *IAES International Journal of Artificial Intelligence*, *9*(1), 65–72. https://doi.org/10.11591/ijai.v9.i1.pp65-72

Ghozy, M., & Pramana, S. (2020). Kajian Penerapan Data Marketplace dalam Penghitungan Indeks Harga Konsumen. *Bachelor Degree Thesis*. https://doi.org/10.13140/RG.2.2.17027.73766

Ibnu, M., & Rachmatullah, C. (2022). The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data. *Teknik Informatika dan Rekayasa Komputer*, *22*(1), 13–24. https://doi.org/10.30812/matrik.v22i1.1803

Idris, A. (2018). *Confusion Matrix*. Medium.Com. https://medium.com/@awabmohammedomer/confusion-matrix-b504b8f8e1d1

Korde, V. (2012). Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. https://doi.org/10.5121/ijaia.2012.3208

Laksana, J., & Purwarianti, A. (2014). Indonesian Twitter Text Authority Classification for Government in Bandung. *International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 129–134. https://doi.org/10.1109/ICAICTA.2014.7005928

Saleh, S. H., Ismail, R., Ibrahim, Z., & Hussin, N. (2019). Issues, Challenges, and Solutions of Big Data in Information Management: An Overview. *International Journal of Academic Research in Business and Social Sciences*, *8*(12). https://doi.org/10.6007/ijarbss/v8-i12/5240

Sitorus, I. (2020). *Support Vector Machine and Kernel Tricks*. Medium.Com. https://medium.com/analytics-vidhya/introduction-to-svm-and-kernel-trick-part-1-theory-d990e2872ace

Srimulyani, W., Pramana, S., & Bustaman, U. (2021). Developing an online shop sampling frame from big data. *Statistical Journal of the IAOS*, *38*, 1483–1490. https://doi.org/10.3233/SJI-220929

Suyanto. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Penerbit Informatika.

Ünal, Y., Sağlam, A., & Kayhan, O. (2019). Improving Classification Performance for an Imbalanced Educational Dataset Example using SMOTE. *European Journal of Science and Technology*, *Special Issue*, 485–489. https://doi.org/10.31590/ejosat.638608

UNECE. (2022). *Machine Learning for Official Statistics*. https://doi.org/10.18356/9789210011143. https://unece.org/media/press/365536

Xu, D., & Wu, S. (2014). An Improved TFIDF Algorithm in Text Classification. *Applied Mechanics and Materials*, *651–653*, 2258–2261. https://doi.org/10.4028/www.scientific.net/AMM.651-653.2258

Yu, W., Sun, Z., Liu, H., Li, Z., & Zheng, Z. (2018). Multi-Level Deep Learning Based e-Commerce Product Categorization. *CEUR Workshop Proceedings*, *2319*.