Yodhi Yuniarthe / Afr.J.Bio.Sc. 6(5)(2024). 10861-10870 **ISSN: 2663-2187**

https://doi.org/ 10.48047/AFJBS.6.5.2024. 10861-10870

African Journal of Biological Sciences

A Comparative Study of Machine Learning Methods for Sentiment Analysis of Lampung Robusta Coffee

Yodhi Yuniarthe[1], Admi Syarif [*2], Sumaryo Gitosaputro[3], Warsito Warsito[4]

[1)]PhD student, Department of Computer Science-Faculty of Mathematics and Natural Sciences, Lampung University, Indonesia
[1)]Department of Informatics Faculty of Computer, Indonesia Mitra University, Indonesia
[2)]Department of Computer Science, Faculty of Mathematics and Natural Science, Lampung University, Lampung, Indonesia
[3)]Department of Agribusiness-Faculty Agriculture, Lampung University, Lampung, Indonesia
[4)]Department of Physics-Faculty of Mathematics and Natural Sciences, Lampung University, Indonesia

* corresponding author: admi.syarif@fmipa.unila.ac.id

(Received July 1, 2019, Revised October 21, 2019, Accepted October 29, 2019, Available online October 29, 2019)

Article History
Volume 6, Issue 5, 2024
Received: 22 May 2024
Accepted: 03 Jun 2024
doi: ***10.48047/AFJBS.6.5.2024.10861-10870***

Abstract

Coffee is known as an international commodity. Two types of coffee beans have a high commercial value, namely Arabica and Robusta. Indonesia is the largest producer of Robusta coffee in the world, and Lampung Province is the best producer of Robusta coffee in Indonesia. People have a lot of opinions about the Robusta coffee from Lampung because of its specialty. This research aims to analyze the sentiment towards Lampung Robusta Coffee based on comments on YouTube. This research aims to analyze the sentiment of Lampung Robusta Coffee on YouTube. This research uses AI methods such as Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbor (KNN) algorithms. We also considered balanced and unbalanced datasets and adopted a data-balancing approach. Overall, the sentiment towards Robusta coffee is mostly positive, with 145 instances (71.4%) expressing positive sentiment and 48 instances (28.6%) expressing negative sentiment. Among the three classification methods, support vector machines achieved the highest accuracy of 82.82% when matching the data with SMOTE, followed by Naive Bayes with 79.54% and K-Nearest Neighbors with 77.38%. The results of this study conclude that the SVM algorithm has the best accuracy on the YouTube comment dataset used in this study.

Keywords: sentiment analysis, support vector machine, naïve bayes, K-nearest neighbor, Lampung robusta coffee

1. Introduction

Indonesia's strategic location in a tropical climate offers advantages and chances for development in agriculture, economics, and other fields. The tropical climate is distinguished by abundant precipitation and ample sunlight, making it conducive to agricultural, horticultural, maritime, forestry, and tourism pursuits. Indonesia is known for its uniqueness and specialties, such as Lampung, which is known for its special species of orchids [1], and crops such as coffee. Indonesia is recognized as the fourth-largest coffee producer globally, with a particular focus on the growth of its plantations. Coffee holds a significant place in people's daily lives and culture worldwide. The Coffee Canephora plant's Robusta coffee variety is integral to the global coffee industry. Coffea canephora L. (Robusta) and Coffea arabica L. (Arabika) are the two most widely popular types of coffee [2]. One of Indonesia's most famous coffee-producing provinces is Lampung. The quality and taste of Lampung Robusta Coffee have attracted the attention of many coffee lovers and stakeholders in the coffee industry [3]. In the digital age, individuals' opinions and perspectives regarding coffee products and brands are readily available on many online forums, social media platforms, and websites.

The rapid advancement of technology undoubtedly significantly impacts the human race. The pace of technological advancements is rapidly accelerating. Artificial intelligence (AI) is a rapidly advancing technology. The advancement of AI has significantly impacted the efficiency of human labor. AI's application is not restricted solely to the telecommunications sector but extends to finance, manufacturing, services, government, and even the agriculture economy. Sentiment analysis is a method used to assess the sentiment expressed in a document or statement and classify it as positive, negative, or neutral. Sentiment analysis applies statistical techniques to investigate, analyze, and extract textual information related to various entities, such as services, products, individuals, phenomena, or subjects. The analytical phase involves assessing multiple forms of textual content such as texts, forums, tweets, or blogs using pre-processed data that includes tokenization, removal of standard terms, elimination of word variations, identification of the root form of words, determination of sentiment, and classification of sentiment [4]. Individuals utilize social media platforms to express their emotions [5]. YouTube is a publicly accessible kind of media. Youtube offers diverse entertainment and informative content, including news, music, and movies. Additionally, YouTube Channels include review menus. A sentimental analysis of Lampung Robusta Coffee is essential to understanding how people respond to and evaluate this product [6]. Sentiment analysis typically falls under supervised learning. This requires annotation of the data [7].

Machine learning (ML) is one of the artificial intelligence (AI) sciences. Large amounts of historical data may be processed by computer systems, and they can use machine learning techniques to find patterns. As a result, the system's predictions based on input data are more accurate. The utilization of machine learning technology in this research has demonstrated its efficacy as a potent instrument for evaluating public opinion and sentiment on a wide-ranging level [6]. Using machine learning techniques, sentiment analysis succeeded worldwide [7]. This research analyzes hospital service sentiment in Lampung using a machine-learning approach [8]. Data from YouTube social media, product reviews, and forum discussions was collected, and then machine learning algorithms were applied to classify sentiment into positive and negative [9]. Three machine learning classification techniques are used in this study: K-nearest Neighbor (K-NN), Naïve Bayes (NB), and Support Vector Machine (SVM). The present research used the SVM technique with Term Frequency-Inverse Document Frequency (TF-IDF) as the strategy for feature extraction. This study diverges from prior research by employing data balancing methods using SMOTE. SVM has the advantage of finding the best hyperplane to separate two classes in feature space and using a Structural Risk Minimization (SRM) strategy for more optimal results [8]. The Word weighting using TF-IDF will be classified into two sentiment values, positive and negative [9]. The feature extraction methods include TF-IDF, SVM, and NB [6]. TF-IDF improves raw term frequency computations by also considering the IDF of each term, which can be characterized by the amount of information each word carries [10]. Once the dataset's preprocessing is finished, the weighting procedure is implemented. This important crucial is converted into a vector representing the word so the system recognizes it [11]. SMOTE is a statistical method that balances data between minority and majority classes [12]. Hande[13] with Amount of data: 7671 Source: YouTube, 59% accuracy for sentiment analysis. 66% accuracy for extensive language detection. Sentiment analysis has been used in previous research, The Sentiment Analysis of Indonesian online travel agent sentiment analysis using machine learning methods [14]. Salma & Silfianti, 2021 [15] with SVM: 76.50%, Naive Bayes: 72.30%, KNN: 59.10%. Previous research on sentiment analysis has led to the development of a large body of knowledge, the results of which have generated knowledge benefits, for example in the business arena, where companies can use sentiment analysis to understand customer views on their products or services. This information can be used to improve the customer experience, identify product problems, or direct marketing strategies, as well as for product development.

This study aims to evaluate the sentiment analysis of Lampung Robusta Coffee. For this study, we collected data from uploads and comments on YouTube. We experimented with the dataset to measure the best performance of machine learning methods and model comparison of three classification algorithms to find the best model. This research will provide valuable insights into people's impressions of Lampung Robusta Coffee. The knowledge gained from this research can be used to make decisions regarding future marketing and development of this coffee product. The programming language uses python as it is one of the most widely used programming languages in scientific computing [16]. The availability of packages like TensorFlow, PyTorch, Scikit-learn, and Keras has contributed to Python's rise in popularity. Python is an open-source language that allows anybody to use and modify major machine-learning libraries and frameworks. It also has great integration with other technologies, like databases, web development, and data analysis.

This research is structured as follows: Section 1 describes the introduction related to Indonesian coffee, artificial intelligence, machine learning, and related research on sentiment analysis, and Section 2 presents the method and workflow. This section discusses steps such as preprocessing, word weighting, holdout validation, training data, support vector machines, naive Bayes, and K-nearest Neighbor algorithms. Section 3 provides a comprehensive explanation of the research results and discussion related to the presentation. This section evaluates the performance

of each classification method. It provides insight into their effectiveness. Section 4 presents conclusions based on the research results.

2. Method

This paper outlines the research procedure illustrated in Figure 1. The first step involves gathering data on the use of YouTube. Following the manual method of assigning a sentiment value to each piece of data, the following step in the data pre-processing process was text cleaning. In this study, word weighting implements TF-IDF, and then data distribution is implemented using holdout validation. The next step uses the sampling method to balance the data. The final process in this study is the performance evaluation of the algorithm using the Confusion Matrix and comparing the three algorithms. The first stage is to identify the problems in the previous research results. We designed the concept model that will be used in the simulation. A preliminary design of the text mining process will be used. The second idea involves manually labeling the comments received after the data has been processed into training and test sets. During the simulation phase, the system is operated to replicate how the algorithm would function in a given environment. The dataset is entered, sentiment labeling is done on the dataset, training is done on the training data, and classification is done using test data to complete the simulation. A comparison of the accuracy of the algorithms employed in this study is the simulation's output. In the final step, the author analyzes the output from the scenario that was run to determine the algorithm's correctness for this study.

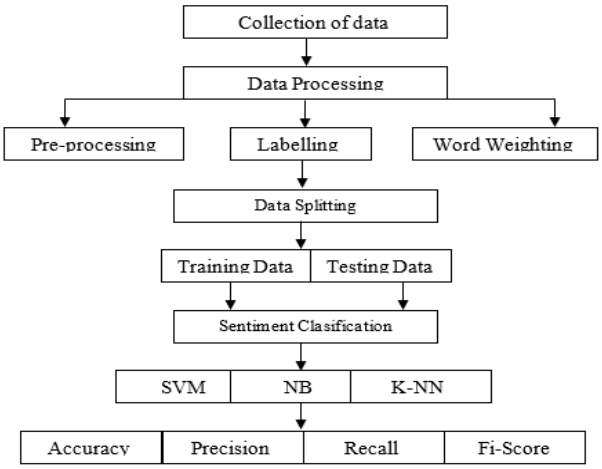

Fig .1. The stages of the research

2.1 Data Collection

Data collection or data crawling is carried out using the YouTube API. The data are then stored in a CSV file.

2.2 Data Processing

Data pre-processing is used after the data-gathering step to cleanse and organize the data for future analysis. Several processes are carried out at this stage: cleaning, case folding, tokenizing, normalization, stopword removal, and stemming [17]. The following is a more detailed explanation of the functions in dataset processing. Cleaning is the process of removing noise or interference from raw text data. At this stage, we will process all research text data in several steps, such as deleting usernames, URLs, hashtag signs (#), mention signs (@), numbers, punctuation marks, HTML characters, and other symbols. Case folding refers to transforming capital characters into lowercase or vice versa in text processing. The main goal is to equalize the representation of words with different capitalization letters, thus facilitating the analysis and processing of text in natural language processing (Natural Language Processing). The tokenization is processed by separating the sentence into words per word, which can be used to analyze sentiment [14]. Tokenization is a method of splitting text into smaller units known as "tokens." The labeling of the dataset was done manually by three labelers [18]. In the labeling process, two main labelers are responsible for determining the sentiment value of each tweet contained in the dataset. The label supporters will also choose the final judgment if the two main labelers have different judgment sentiments. In this study, Polaris is used as a measure of sentiment. Positive polarity is used for data containing positive sentiments, while negative polarity is used for data containing negative sentiments. One hundred forty-five data, or 71.4%, have a positive sentiment, and 58 data, or 28.6%, have a negative sentiment. The manual labeling's outcomes are displayed in Table 1.

Table 1. Manual Labelling Results

Comment	Sentiment
Remember when I was in Lampung...picked my ongsreng...mashed it myself...wow, it's delicious.	Positive
Lampung robusta now tastes bitter like regular coffee	Negative

We utilize the Term Frequency-Inverse Document Frequency (TF-IDF) methodology in the present study. The TF-IDF approach quantifies the significance of a word or phrase within a dataset. This weighting process involves two main components: Term Frequency (TF) and Inverse Document Frequency (IDF) [19]. The term frequency (TF) measures how often certain words appear in a document. The more often the word appears, the higher the weight. Inverse Document Frequency (IDF) measures the word's uniqueness level in the entire dataset. Using the TF-IDF method, words frequently appearing in the document but rarely appearing in other documents will have a high weight [20]. Conversely, words that occur often in the entire dataset will have a low weight due to their lack of uniqueness. The result of this word weighting will be a numerical representation of each document in the dataset [21].

2.3 Data Splitting

As shown in Figure 1, we split the data distribution used in this study into training and test data. Holdout validation is part of the data distribution procedure. The data distribution is done by holdout validation. The division of data training and data testing into five scenarios: Scenario 1 is 50% data training and 50% data testing; Scenario 2 is 55% data training and 45% data testing; Scenario 3 is 60% data training and 40% data testing; Scenario 4 is 65% data training and 35% data testing; and Scenario 5 is 70% data training and 30% data testing.

2.4 Sentiment Classification

This study's sentiment classification stages use three algorithms: SVM, Naïve Bayes, and K-NN, three different machine learning algorithms. The dataset was divided into training and test data in the early stages. The SMOTE technique addresses class imbalances through Synthetic Minority Over-sampling by re-sampling the minority class sample. The library used to balance data is imbalanced-learn [22]. The results were obtained by implementing the SMOTE into the balanced dataset. The training data helps train the classification algorithm, while the test data helps Analyze the modified algorithm's performance. The classification results are in the form of positive and negative sentiment predictions—the evaluation utilized metrics such as the Confusion Matrix, Accuracy, Precision, Recall, and F1-Score.

2.4.1 Support Vector Machine (SVM)

The SVM is a learning algorithm that generates hypothesis spaces using linear functions in high-dimensional feature spaces. We conduct this training using learning algorithms based on the concepts of optimization theory The power of an SVM lies in its ability to learn data classification patterns with balanced accuracy and reproducibility. SVM has become a widely used tool for classification, with a high level of versatility that spans multiple data science scenarios [23]. SVM can be classified into two primary categories: linear SVM and nonlinear SVM. A hyperplane with a soft margin is used in linear support vector machines (SVM) to split data into related classes. In addition, non-linear SVM maps data into a higher-dimensional space for improved separation by using the kernel method [24].

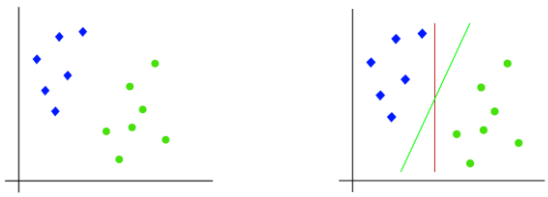

Fig .2. Linear SVM [25]

2.4.2 Naïve Bayes (NB)

Maximum likelihood estimate is used in Naive Bayes classification to group samples into the most likely classes [26]. In this context, if we have an input vector, X, comprising features and a corresponding class label, Y, the notation P(Y|X) represents naive Bayes. This notation represents the posterior probability for Y, which is the likelihood of detecting class label Y given features X. The process of categorization also takes into account the initial possibility, P(Y), which represents the prior probability. Using the information from the training data, the task during training is computing the posterior probabilities (P(Y|X)) for each combination of X and Y [27].

2.4.3 K-Nearest Neighbor (K-NN)

The k-nearest neighbor (K-NN) algorithm is a supervised machine learning algorithm for classification tasks. The k-NN method is a predictive model that assigns labels to data points based on their proximity to the nearest neighbors. The three generally used distance metrics are the Euclidean Distance metrics commonly used in mathematics and computer science, including the Manhattan and Minkowski Distance. After calculating the distance, look for K-NN near the new data. The simple principle of this method is data that will predict whether it belongs to the positive or negative class [28].

2.5 Evaluation

The evaluation in this work involves comparing sentiment analysis performance utilizing three algorithms: SVM, NB, and K-NN with distance metrics such as Manhattan Distance and Minkowsky Distance. Upon computing the distance, search for K-NN near the new data. Tests on each dataset aim to determine Accuracy, Precision, Recall, and F1-Score value changes. After the classification process using SVM, NB, and K-NN, comparisons were made between the classification results of the three algorithms. This evaluation generates the Confusion Matrix and Classification Report, which include metrics such as accuracy, precision, recall, and f1-score. The Confusion Matrix provides information regarding the expected classification results of the algorithm and the actual information labeled by humans. The true negative (TN) value represents the accurate detection of harmful data, whereas the false positive (FP) value refers to positive data being incorrectly identified as harmful. TP refers to positive data that is accurately identified, while FP refers to harmful data that is mistakenly recognized as positive. Classification system performance is generally calculated using data from The confusion matrix, as shown in Table 2 below.

Table 2. Confusion Matrix

Actual	Prediction	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

The actual negative (TN) value accurately detects harmful data, whereas the false negative (FN) value refers to positive data being incorrectly identified as negative. True positive (TP) refers to positive data that is accurately identified, while false positive (FP) refers to negative data that is mistakenly recognized as positive. The Confusion Matrix table is used to measure the performance of a classification method by calculating the value of accuracy, precision, recall, and f1-score [29].

a. Accuracy

Accuracy predicts true positive and negative data from the entire dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

b. Precision

Precision is an objective measure of the level of correctness in the results generated by a model. Precision is the ratio of accurate optimistic predictions to the total positive predictions made.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

c. Recall

Recall measures model completeness. The confusion matrix gives insights into expected categorization outcomes for labeled data.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

d. F1-Score

The F1-Score compares the average value obtained from recall and precision.

$$F1\ Score = 2 \times \frac{(recall \times precision)}{(recall + precision)} \quad (4)$$

3. Results and Discussion

In this study, we have done intensive numerical experiments. The dataset is taken from YouTube comments about coffee robusta lampung form with text attributes from August 2023 to September 2023. First, we got 942 datasets, and then after filtering and labeling, we had 203 (145 positives, 58 negatives). The data split into data training and data testing by using five different scenarios as follows:

Table 3. Composition of training and testing

Scenario	Training	Testing
1	50%	50%
2	55%	45%
3	60%	40%
4	65%	35%
5	70%	30%

We also considered the balanced and unbalanced datasets for this experiment and adopted a data-balancing approach. The results of research on imbalance and balanced data with five scenarios, according to Table 3, are as follows:

Table 4. Scenario with imbalanced and balanced datasets

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5			Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
Inbalanced	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Balanced	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
	50%	50%	55%	45%	60%	40%	65%	35%	70%	30%		50%	50%	55%	45%	60%	40%	65%	35%	70%	30%
Positive	72	73	79	66	86	59	94	51	101	41	Positive	72	73	79	66	86	59	94	51	101	44
Negative	29	29	32	26	35	23	37	21	44	17	Negative	72	29	79	26	86	23	94	21	101	17
Total	101	102	111	92	121	82	131	72	142	61	Total	144	102	158	92	172	82	188	72	202	61

We used three methods for training and testing: SVM, NB, and K-NN algorithms. The oversampling method used to handle a class imbalance in classification problems uses SMOTE, which aims to increase the number of samples in the minority class by creating new synthetic samples based on existing samples in that class. The data selection in this study can be seen in Table 4, where the composition between training and testing data has been determined. The Python 3 programming language was used as a development tool for the data-splitting process, and the sci-kit-learn library was used to split the number of datasets. For classification modeling, the Sklearn package was used with the given dataset, specifically the Support Vector Classification (SVC) algorithm. Data can be classified into two categories: training and testing data. Then, testing uses several kinds of support vector machine kernels.

The four kernels used in this study are linear, radial basis function (RBF), polynomial, and sigmoid. The objective is to determine the kernel that achieves the maximum level of accuracy. The accuracy outcomes obtained from the SVM in the 2nd scenario are accuracy of 79.35%, precision of 78.31%, recall of 98.48%, and f1-score of 87.25%. The Confusion Matrix for the best scheme for the imbalanced dataset for the 2nd scenario and the balanced dataset for the 5th scenario. The Naïve Bayes using data balancing techniques (SMOTE) can provide the best accuracy results of 80.56%. Precision is 81.36%, recall is 94.12%, and F1-score is 87.27% in the 4th scenario. Whereas for the imbalanced dataset, the accuracy results with the naïve Bayes in the 3rd scenario are 70.73%, precision 72.15%, recall 96.61%, and f1-score 82.61%. The Confusion Matrix for the best scheme for the imbalanced dataset for the 3rd scenario and the balanced dataset for the 4th scenario. For the K-NN classification in this study, we are using a parameter of five neighbors, with weight based on distance and p equal to 1, to ensure the highest level of accuracy.

The testing results were based on five data-sharing scenarios with imbalanced and balanced datasets. Using data balancing techniques (SMOTE), the K-NN can provide the best accuracy results of 79.41%. Precision is 84.21%,

recall is 87.69%, and F1-score is 85.91% in the 1st scenario. Whereas for the imbalanced dataset, the accuracy results with the K-NN in the 2nd scenario are accuracy is 73.91%, precision is 75.61%, recall is 93.94%, and f1-score is 83.78%.

In this study, the Confusion Matrix calculates various metrics such as Accuracy, Precision, Recall, and F1-Score. A comparison of the three algorithms takes accuracy as a reference. The results of the accuracy comparison of the Support Vector Machine, Naïve Bayes, and K-nearest Neighbor algorithms are in Table 6. The algorithm that gives the best accuracy results is the SVM using the SMOTE of 86.89% in the 5th scenario, Naïve Bayes of 80.56% in the 2nd scenario, and K-NN of 79.41% in the 1st scenario. SMOTE improves accuracy results for imbalanced datasets using the three algorithms. The highest balanced average value is SVM, with a score of 83.91, and imbalanced, with a value of 77.71%. According to Table 6, for the performance of algorithmic methods on imbalanced and balanced datasets, the distribution of the comparison result data is as follows: In the imbalanced dataset scenario, SVM tends to provide overall better results than NB and K-NN. SVM has the highest average score for all evaluation metrics. In the balanced dataset scenario, K-NN performed better than NB and SVM, with the highest average values for most evaluation metrics.

The results of the performance comparison between the algorithmic methods are as follows: On the unbalanced dataset, SVM has the highest average value for all evaluation metrics, followed by K-NN, and NB has the lowest performance. SVM has the best performance with the highest average value for all evaluation metrics (accuracy, precision, recall, and F1 score) on both unbalanced and balanced datasets.

Figure 3 shows that the SVM algorithm has the highest accuracy on balanced data with a value of 82.82%, followed by NB with 79.54% and K-NN with an accuracy of 77.38%.

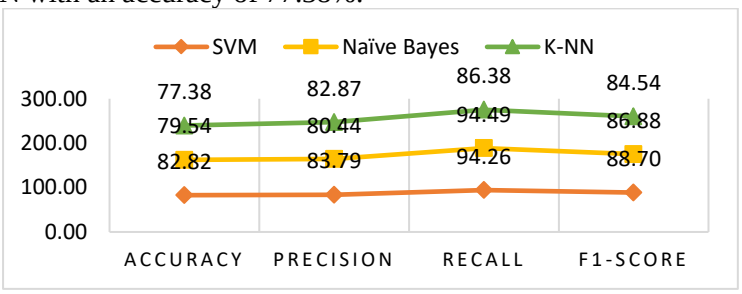

Fig .3. Graph of balanced result

Table 5. The comparison of the three method algorithms (Percentage)

Scenario		SVM		*Naïve Bayes*		K-NN	
		Unbalanced	*Balanced*	*Unbalanced*	*Balanced*	*Unbalanced*	*Balanced*
1	Accuracy	77,45	78,43	69,61	80,39	73,91	79,41
	Precision	76,6	80,72	71	80,46	75,61	84,21
	Recall	98,63	91,78	97,26	95,89	93,94	87,69
	F1-Score	86,23	85,9	82,08	87,5	83,78	85,91
2	Accuracy	79,35	84,78	70,65	77,17	78,26	78,26
	Precision	78,31	86,11	71,91	77,78	81,94	81,94
	Recall	98,48	93,94	96,97	95,45	89,39	89,39
	F1-Score	87,25	89,86	82,58	85,71	85,51	85,51
3	Accuracy	78,05	79,27	70,73	79,27	73,17	74,39
	Precision	77,33	82,81	72,15	81,82	75,34	80,65
	Recall	98,31	89,83	96,61	91,53	93,22	84,75
	F1-Score	86,57	86,18	82,61	86,4	83,33	82,64
4	Accuracy	76,39	84,72	69,44	80,56	72,22	77,78
	Precision	75,76	83,33	71,01	81,36	74,6	81,82
	Recall	98,04	98,04	96,08	94,12	92,16	88,24
	F1-Score	85,47	90,09	81,67	87,27	82,46	84,91
5	Accuracy	77,05	86,89	70,49	80,33	73,77	77,05
	Precision	76,79	86	72,41	80,77	76,92	85,71
	Recall	97,73	97,73	95,45	95,45	90,91	81,82
	F1-Score	86	91,49	82,35	87,5	83,33	83,72
Average	Accuracy	77,66%	82.82%	70,18%	79,54%	73,12%	77,38%
	Precision	76,96%	83,79%	71,70%	80,44%	75,44%	82.87%
	Recall	93,96%	94,07%	94,57%	94,49%	92,68%	86,38%
	F1-Score	86,30%	88,70%	82,26%	86,88%	83,17%	84,54%

The comparison results of the accuracy, precision, recall, and F1 scores vary greatly, as shown in Table 5. These values show the computational level of the research using three algorithms. There is a variation of values between unbalanced and balanced with the five scenarios used. From these results, the SVM algorithm has the highest accuracy, precision, recall, and F1 score values.

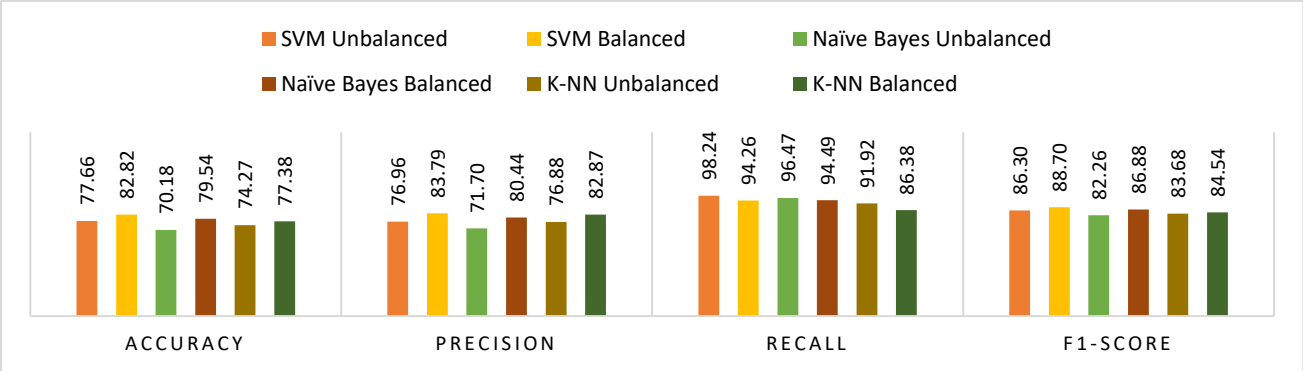

Fig .4. Graph of tree algorithm result

Figure 4 shows the results of the algorithm comparison, the SVM algorithm with imbalanced data performed very well on recall and F1-Score metrics; however, the SVM with balanced data performed better on accuracy and precision. When faced with imbalanced data, Naive Bayes often performs worse than the others across all metrics; however, K-NN performance varies less significantly when faced with balanced or unbalanced data. In general, using balanced data enhances the algorithms' overall performance across a range of criteria, particularly for SVM.

4. Conclusion

This research analyzed Lampung Robusta Coffee's sentiment using a Support Vector Machine, Naïve Bayes, and K-nearest Neighbor. The dataset comprised 203 instances collected from YouTube. Overall, the opinion of Robusta Coffee was generally positive. Lampung Robusta Coffee had a predominance of positive sentiment. One hundred forty-five had a positive sentiment of 71.4%, and 48 had a negative sentiment of 28.6%. The Support Vector Machine has shown the best accuracy results method by balancing SMOTE data at 82.82%, then in Naïve Bayes at 79.54% and K-nearest Neighbors at 77.38%, so it can be concluded that SVM has the best accuracy on the YouTube comment dataset used in this research. Sentiment analysis offers the chance to mine rich and diverse textual data for insightful information that can enhance decision-making in a variety of domains.

Acknowledgment

The author expresses his profound appreciation to Universitas Mitra Indonesia (UMITRA), which has funded internal research grants with grant numbers No. 04/UMITRA/HIBAH/FKOM/2024.

References

[1] Mahfut*,Tundjung Tripeni Handayani, Sri Wahyuningsih, and Sukimin, "Identification of Dendrobium (Orchidaceae) in Liwa Botanical Garden Based on Leaf Morphological Characters," *Journal of Tropical Biodiversity and Biotechnology (JTBB).*, vol. 06, no. 01, pp. 1–6, 2021, doi: 10.22146/jtbb.59423.

[2] M. Jeszka-Skowron, R. Frankowski, and A. Zgo, "Comparison of methylxanthines, trigonelline, nicotinic acid, and nicotinamide contents in brews of green and processed Arabica and Robusta coffee beans – Influence of steaming, decaffeination and roasting processes," *Food Sci. Technol (LWT).*, vol. 125, no. 5, pp. 1–9, 2020, doi: 10.1016/j.lwt.2020.109344.

[3] H. R. Santosa, C. Cucu Suherman, and S. Rosniawaty, "Responses Growth of Coffee Plants Robusta (Coffea robusta L.) Aluminum in Ground Reclamation Ancient Coal Mines Vegetation Sengon (Period)," *Agrikultura*, vol. 27, no. 3, pp. 124–131, 2016, doi: 10.24198/agrikultura.v27i3.10871.

[4] R. Rasenda, H. Lubis, and R. Ridwan, "Implementation of K-NN in Usury Sentiment Analysis on Bank Interest Based on Twitter Data," *J. Media Inform. Budidarma*, vol. 4, no. 2, pp. 369–376, 2020, doi: 10.30865/mib.v4i2.2051.

[5] A. M. Rahat, A. Kahir, A. Kaisar, and M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *Proceedings of the SMART–2019, IEEE Conference*, University of Exeter, Ed., Moradabad, India: Proceedings of the SMART–2019, IEEE, 2019, pp. 266–270. doi:

10.1109/SMART46866.2019.9117512.

[6] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no. 1, pp. 45–55, 2021, doi: 10.26555/jifo.v15i1.a20111.

[7] M. Liebenlito, N. Inayah, E. Choerunnisa, T. E. Sutanto, and S. Inna, "Active Learning on Indonesian Twitter Sentiment Analysis Using Uncertainty Sampling," *J. Appl. Data Sci.*, vol. 5, no. 1, pp. 114–121, 2024, doi: DOI: https://doi.org/10.47738/jads.v5i1.144.

[8] K. Munawaroh and A. Alamsyah, "Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter," *J. Adv. Inf. Syst. Technol.*, vol. 4, no. 2, pp. 113–125, 2023, doi: 10.15294/jaist.v4i2.59493.

[9] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *Indonesian J. Comput. Cybern. Syst (IJCCS).*, vol. 15, no. 2, pp. 121–130, 2021, doi: 10.22146/ijccs.65176.

[10] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27. Elsevier, Finland, pp. 16–32, 2018. doi: 10.1016/j.cosrev.2017.10.002.

[11] Y. FU and Y. Yu, "Research on Text Representation Method Based on Improved TF-IDF," in *Journal of Physics: Conference Series*, Y. F. and Y. Yu, Ed., Wuhan China: IOP Publishing, 2020, pp. 1–8. doi: 10.1088/1742-6596/1486/7/072032.

[12] Y. Man, M. Ng, K. H. Pham, M. Luengo-Oroz, Y. Man, and M. Ng, "Exploring YouTube ' s Recommendation System in the Context of COVID-19 Vaccines : Computational and Comparative Analysis of Video Trajectories Corresponding Author :," *J. Med. Internet Res.*, vol. 25, no. e49061, pp. 1–16, 2023, doi: 10.2196/49061.

[13] A. Hande, R. Priyadharshini, and B. R. Chakravarthi, "KanCMD : Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection," in *Proceedings of the Third Workshop on Computational Modeling of People's Opinions*, E. D. Malvina Nissim, Viviana Patti, Barbara Plank, Ed., Barcelona, Spain: Association for Computational Linguistics, 2020, pp. 54–63. [Online]. Available: https://aclanthology.org/2020.peoples-1.6.pdf

[14] A. D. Poernomo and S. Suharjito, "Indonesian online travel agent sentiment analysis using machine learning methods," *Indonesian Journal of Electrical Engineering and Computer Science*., vol. 14, no. 1, pp. 113–117, 2019, doi: 10.11591/ijeecs.v14.i1.pp113-117.

[15] A. Salma and W. Silfianti, "Sentiment Analysis of User Reviews on COVID-19 Information Applications Using Naive Bayes Classifier, Support Vector Machine, and K-Nearest Neighbor," *Int. Res. J. Adv. Eng. Sci.*, vol. 6, no. 4, pp. 158–162, 2021.

[16] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," *J. ofMachine Learn. Res.*, vol. 12, no. 3, pp. 2825–2830, 2011.

[17] H. Dag, "The impact of text preprocessing on the prediction of review ratings," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1405–1421, 2020, doi: 10.3906/elk-1907-46.

[18] Y. Sibaroni and A. Tama, "Labeling Analysis in the Classification of Product Review Sentiments by using Multinomial Naive Bayes Algorithm Labeling Analysis in the Classification of Product Review," in *The 2nd International Conference on Data and Information Science*, W.-X. Z. e Zhi-Qiang Jiang, Wen-Jie Xie, Ed., Bandung, Indonesia: IOP Publishing Ltd, 2019, pp. 1–12. doi: 10.1088/1742-6596/1192/1/012036.

[19] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text .11591/eei.v10i5.3157.

[20] A. M. Pravina, "Sentiment Analysis of Delivery Service Opinions on Twitter Documents using K-Nearest Neighbor," *Jurnal Tek. Inform. dan Sist. Informasi (JATISI)*, vol. 9, no. 2, pp. 996–1012, 2022, doi: 10.35957/jatisi.v9i2.1899.

[21] A. Nugroho, K. Mandara, and Ricky Risnantoyo, "Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm," *Journal Informatics Telecommun. Eng (JITE).*, vol. 4, no. 1, pp. 86–96, 2020, doi: 10.31289/jite.v4i1.3798.

[22] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res. 16*, vol. 16, no. 3, pp. 321–357, 2002, doi: https://doi.org/10.1613/jair.953.

[23] D. A. Pisner and D. M. Schnyer, *Support vector machine*. United States: Elsevier Inc., 2020. doi: 10.1016/B978-0-12-815739-8.00006-7.

[24] FR Lumbanraja, E Fitri, Ardiansyah, A Junaidi, and R. Prabowo "Abstract Classification Using Support Vector

Machine Algorithm (Case Study: Abstract in a Computer Science Journal)," *J. Phys. Conf. Ser.*, pp. 1–13, 2021, doi: 10.1088/1742-6596/1751/1/012042.

[25] A. Syarif, O. D. Riana, D. A. Shofiana, and A. Junaidi, "A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification : Decision Tree, Support Vector Machine, and Naive Bayes," *International Journal of Advanced Computer Science and Applications* (*IJACSA)*, vol. 14, no. 10, pp. 597–603, 2023, doi: 10.14569/issn.2156-5570.

[26] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *J. Adv. Signal Process (EURASIP).*, vol. 30, no. 3, pp. 1–12, 2021, doi: 10.1186/s13634-021-00742-6.

[27] D. Ajeng 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," *2018 6th Int. Conf. Cyber IT Serv. Manag.*, no. Citsm 2018, pp. 1–6, 2023, doi: 10.1109/CITSM.2018.8674352.

[28] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in the classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp. 86–91, 2021, doi: 10.21831/jeatech.v2i2.42777.

[29] K. Muludi, M. S. Akbar, D. A. Shofiana, and A. Syarif, "Sentiment Analysis Of Energy Independence Tweets Using Simple Recurrent Neural Network," *Indonesian Journal Computer Cybern. System (IJCCS).*, vol. 15, no. 4, pp. 339–348, 2021, doi 10.22146/ijccs.66016.