Predictions of Indonesia economic phenomena based on online news using Random Forest

Fitri Khairani[1)*], Anang Kurnia[2)], Muhammad Nur Aidi[3)], Setia Pramana[4)]
[1)2)3)] IPB University, Indonesia, [4)] Politeknik, Statistika STIS, Jakarta, Indonesia
[1)]fitrikhairanil@apps.ipb.ac.id, [2)]anangk@apps.ipb.ac.id, [3)]muhammadai@apps.ipb.ac.id,[4)]
setia.pramana@stis.ac.id

Submitted : Apr 21, 2022 | **Accepted** : Apr 24, 2022 | **Published** : Apr 27, 2022

Abstract: Economic growth in the first quarter of 2021 based on YoY (Year on Year) is around -0.74%. This figure caused the Indonesian economy to recession after contracting four times since the second quarter of 2020. With positive and negative growth in the value of GDP for each category based on the business sector each quarter, can do future economic growth modelling. The prediction results can be used as an early warning for the government on factors that can maximize and factors that must improve. This study aims to predict the state of economic growth in the next quarter using Random Forest classification. Random Forest combines tree classification and bagging by resampling the data, which reduces the variance of the final model, which is for low variance overfitting. The data used in this study was scrapped from January 2021 to March 2021 on 5 Indonesian online news portals, namely Kompas, Antara, Okezone, Detik, and Bisnis. The independent variable is online news based on GDP category. The dependent variable results from data labelling on each news, up or down, carried out by the Directorate of Balance Sheet of BPS. Based on the calculations with cross-validation of 10, the modelling results obtained 96.51% accuracy, 97% precision, and 97% recall. The random forest method is good for predicting economic growth in the next quarter, namely the second quarter of 2021. Incorrectly predicted only three categories of GDP were: the construction category, the transportation and warehousing category, and the company service category.

Keywords: Classification, Economy, GDP, Online News, Prediction, Random Forest

INTRODUCTION

Gross Domestic Product (GDP) is the added value generated from all existing business units in a country or the total value of final goods and services produced by all economic units. According to data from the Central Statistics Agency (BPS), economic growth in the first quarter of 2021 based on YoY (Year on Year) is around -0.74%. This figure causes the Indonesian economy to remain in a recession after experiencing contractions four times since the second quarter of 2020. With positive and negative growth in the value of GDP based on business fields, it is necessary to predict future economic growth. Predicting future economic growth is done as an early warning to the government on factors that can maximize and factors that must improve for each category of GDP in the business field.

According to data from the Ministry of Communication and Information (Kominfo), in 2018. Forty-three thousand online news portals, 100 of which have been verified by the press council. Much online news makes information can be obtained more quickly and variously from various sources. The trend of specific topics discussed a lot would be published in online news along with the opinions of specific experts and experts. The number of online news that is published every day is enormous. Online news is an essential source of data to provide information on events, so in this study, economic growth predictions in each category of GDP are carried out.

Several studies related to news text classification have been carried out, and the methods used also vary, such as support vector machines, random forests, naive Bayes, and others. Classification refers to identifying an observation in a specific class that has been determined. The classification of news related to economic phenomena aims to predict whether the next quarter's economic growth will rise or fall, which can later be used as an initial picture in determining policy.

*name of corresponding author

The analysis used the random forest method. The random forest method combines the tree classification method and bootstrap aggregating (bagging). The basic idea of the random forest method is that a model consists of several tree classification models by resampling the data. This process produces a collection of single trees of different sizes and shapes. The expected result is that a collection of single trees has a small correlation between trees. This small correlation causes the estimated random forest yield variance to be smaller than the estimated range of bagging results, thereby reducing overfitting. Each tree will display the classification results, and Random Forest selects the results that appear the most using aggregation. This study produces economic growth predictions for each category of GDP and the accuracy of the prediction results.

LITERATURE REVIEW

Previous research on text classification of news articles in Indonesian was carried out by (Wongso et.al, 2017) using the Multinomial Naive Bayes method. This study uses Hasir Scrapping data from CNN Indonesia Online News with 5,000 documents. These documents consist of 1,000 documents for each category: economy, health, sports, politics, and technology. Documents are randomly divided by 80:20 ratios for training and objectives. Based on the test results, the combination of TF-IDF and Multinomial Naïve Bayes Classifier provides precision results of 0.9841519 and the recall of 0.9840000.

The following study (Barua et.al, 2021) discussed the multi-class sports news categorization using Machina Learning in 2021. The data used was Bengali news corpus (called BNEC) consisting of 43306 news documents with 202830 unique words in several classes: Cricket, soccer, tennis, and athletics. The experimental results of the dataset test showed that the Support Vector Classifier (SVC) with the Unigram + Trigram + Trigram feature room obtained an F1 score weighing 97.60%.

A study (Suleymanov et.al, 2018) discussed the classification of Azerbaijan news articles. This study uses using SVM, NB, and Ann. The study results with the Ann + MLP method obtained maximum accuracy (89.1%). The study uses Random Forest for the classification of big data (Lakshmanaprabu et al., 2019). The size reduction is performed using the Map Reduce framework. Experimental results show that the performance of the proposed RF is high as compared to traditional RF with respect to the accuracy, precision, and recall. Similarly, a study (Parida et al., 2021) discussed news text classification is investigated using RF and NB with TFIDF features. Results prove the superiority of RF over NB for news categorization.

The following research was conducted by (Dhar et.al, 2021), which discussed the categorization of Bengali news titles using The optimized machine learning pipeline. The data used is news headlines from various popular news portals in Bengali. Data is collected in three categories: health, sports, and technology. The size of the data set is 1000x3. The results obtained by classifiers are NB for three categories, namely the highest F1 score of 81%.

METHOD

Text Mining

Text mining is extracting patterned information from large amounts of text data, such as word documents, news, PDFs, text quotes, or even text mining—SMS (tweet). Text mining is divided into two stages, first starting with changing unstructured text data to structured data, followed by extracting the researched information from structured text data. Text mining can be classified (classifier) or just by looking at the frequency (word cloud) (Pathak, 2014).

Based on the irregularity of the text data structure, the text mining process first performs the pre-processing stage:

1. Case Folding

Case folding is the stage of changing all capital letters in the document into lowercase characters other than letters that will remove (Manning et al., 2009) .

Table 1 Example of Case Folding Steps

Before	After
Financial management is a piece of essential knowledge.	financial management is a piece of essential knowledge.

2. Cleaning Data

A process of cleaning words in a document by removing punctuation marks such as commas (,), periods (.), semicolons (;), colons (:), mentions, and others that are less important to reduce noise (Manning et al., 2009) .

Table 2 Example of Cleaning Data Steps

*name of corresponding author

Before	After
financial management is a piece of essential knowledge.	financial management is a piece of essential knowledge

3. Filtering

Filtering is the process of deleting less important words, such as pronouns, conjunctions, adverbs, and others, by using a stopword, a list of words deleted in the document (Manning et al., 2009).

Table 3 Example of Filtering Steps

Before	After
financial management is a piece of essential knowledge	financial management essential knowledge

4. Tokenizing

Tokenization is cutting a document into tiny fractions, which can be in the form of chapters, sub-chapters, paragraphs, sentences, and words (tokens) (Manning et al., 2009).

Table 4 Example of Tokenizing Steps

Before	After
financial management essential knowledge	[‘financial’,’management’,’essential’,’ knowledge’]

Word Weighting (TF-IDF)

TF-IDF is a metric that multiplies the two quantities tf and idf. Here, tf provides a direct estimation of the occurrence probabilityof a term when it is normalized bythe total frequencyin the document, or the document collection, depending on the scope of the calculation. Note that the normalization factor is common for all the terms in the scope, and thus can be omitted. On the other hand, idf can be interpreted as ‘the amount of information’ in conventional information theory (Brookes, 1972; Wong & Yao, 1992), given as the log of the inverse probability (Cover & Thomas, 1991). Stages of weighting with TF-IDF :

Table 5 Table term matriks

	Term	Term	Term	....	Term
Doc 1	$f_{1,1}$	$f_{2,1}$	$f_{3,1}$	....	$f_{i,1}$
Doc 2	$f_{1,2}$	$f_{2,2}$	$f_{3,2}$	....	$f_{i,2}$
....	....	....	....	....	....
Doc j	$f_{1,j}$	$f_{2,1}$	$f_{3,j}$	....	$f_{i,j}$

1. Calculate term frequency $tf_{t,d}$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

2. Calculate document frequency (df)
3. Calculate bobot inverse document frequency (idf)

$$idf = \log \frac{N}{df_t} \tag{2}$$

4. Calculate weight value TF-IDF

$$W_{i,j} = tf_{i,j} \times \text{idf}$$ (3

Description:

$tf_{i,j}$ = Frequency term

*name of corresponding author

$n_{i,j}$ = Number of words i in document j
df = Number of frequency of documents containing
N = Total number of documents
$W_{i,j}$ = TF-IDF weight

Random Forest

Random Forest is a classification method (supervised learning) that has a grouping on its dependent variable. Random Forest combines the technique of tree classification and bootstrap aggregating (bagging). Leo Breiman developed a random Forest. The Random Forest consists of several tree classification models by resampling the data. Every tree will display the classification results, which Random Forest then selects; most appear by aggregation (Breiman, 2001).

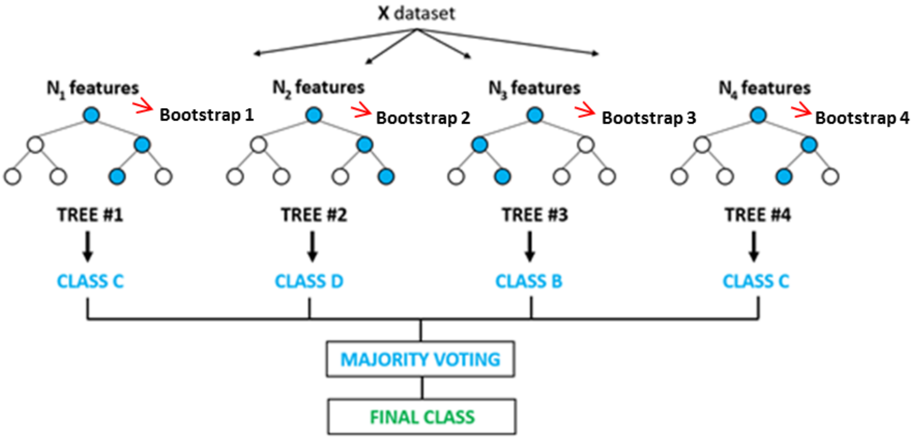

Figure 1 Steps of Random Forest

Based on Figure 1, the concept of the stages of preparation and estimation using random forest is

1. The initial step is to bootstrap up to the number l, which is to draw samples randomly from the independent variable's training data owned by the return size n.
2. Using the examples in each bootstrap to build the tree until it reaches the maximum size. Arrange the tree based on the bootstrap data. Each separation process selects m < p as the independent variable, and the best separation is done in the random sub-setting stage.
3. Repeat steps 1-2 l times to form a forest consisting of the tree.
4. Make predictions based on the prediction results on each result dataset Bootstrapping using majority vote for classification cases

In Random Forest modelling, every time a tree is formed, the independent variable used to perform the separation is not all the variables involved but only part of the random selection results. This process generates a collection of single trees of different sizes and shapes, and the expected result is a collection of single trees that have little correlation between trees. Correlation This small size resulted in a slight variance of the estimated random forest results (Hastie et.al, 2008) and smaller than the estimated variance of the bagging results (Zhu, 2008) to reduce overfitting.

The random forest classification error is estimated through the OOB (Out of Bag) error obtained by (Breiman, 2001):

1. Make predictions on each OOB data, namely data not contained in the bootstrap example, in the appropriate tree.
2. On average, each of the original data observations will be OOB for about 36% of the trees. Therefore, in step l, each initial data observation is predicted to be about a third of the number of trees. Suppose a is an observation from the original data cluster. The random forest prediction result combines the prediction results each time a becomes OOB data.
3. OOB error is calculated from the proportion of errors in random forest prediction results from all over original data cluster observations.

Model Evaluation and Validation

*name of corresponding author

Classification analysis is expected that can classify all data correctly. However, sometimes there are some errors in classification. Validation and evaluation are needed to assess whether the model is correct or not. To avoid errors following the application of the model to new unknown data, namely overfitting, it is necessary to assess the model. Dividing the dataset into training and test data is an approach to avoid overfitting. Training data is used to build the model and test data to validate the built model (Mohammed et al., 2017).

The matrix in Table 1 serves as a basis for comparison in identifying the optimal method based on the data that has been analyzed. Model validation can be done by comparing new built model size by using confusion matrix (Luque et al., 2019). The confusion matrix is used to present the performance of a classifier information. (Bramer, 2007).

Several calculations that determine the performance of model predictions in text classification are accuracy, precision, sensitivity, and f1 score. These four calculations are the confusion matrix shown in Table 6.

Table 6 Confusion Matrix for Two Class

Actual Class	Prediction Class	
	+	-
+	True Positives (TP)	False Negatives FN)
-	False Positives (FP)	True Negatives (TN)

True Positive (TP) is an outcome where the model correctly predicts the positive class.
True Negative (TN) is an outcome where the model correctly predicts the negative class.
False Positive (FP) is an outcome where the model incorrectly predicts the positive class.
False Negative (FN) is an outcome where the model incorrectly predicts the negative class.

Accuracy is a measure of how close the measurement result is to the actual value, according to (Sarkar, 2019). Accuracy is defined as the overall accuracy or the proportion of correct predictions of the model. The calculation of accuracy is written as in Equation 4.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \qquad (4)$$

Precision is the ratio of the correct positive predictions to the total number of positive predictions. The calculation of precision is written as in Equation 5.

$$\text{Precision} = \frac{\text{TN}}{\text{TN+FP}} \qquad (5)$$

Recall and hit rate, coverage, or sensitivity compare True Positive (TP) and the positive amount of data. The recall calculation is written as in Equation 6.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (6)$$

The F1 score is another measurement of accuracy, calculated by taking the harmonic average of precision and recall. The calculation of the F1 score is written as in Equation 7.

$$\text{Skor F1} = \frac{1}{2}\left(\frac{1}{\text{precision}} + \frac{1}{\text{recall}}\right) \qquad (7)$$

The classification model has good precision and recall value if the F1-Score value is greater or closer to 1.

Data Description

The data used in this study uses primary data by scrapping using python software from January 2021 to March 2021 on several Indonesian online news portals, which are grouped by category and subcategory of business fields. List of online news portals that will be scrapping: (1) Kompas, (2) Antara, (3) Okezone, (4) Detik, (5) Bisnis.

The five online news portals were chosen, including the websites with the most visitors in Indonesia, as one of the largest news agencies in Asia and supported by global news networks. Because of economic phenomena, business news portals and the five online news portals have been verified by the press council.

Table 7 Data Description

Variable	Definition	Scale
Y	Data labeling on each news, up or down	Nominal
X	Online News	Text

Stage Of Analysis

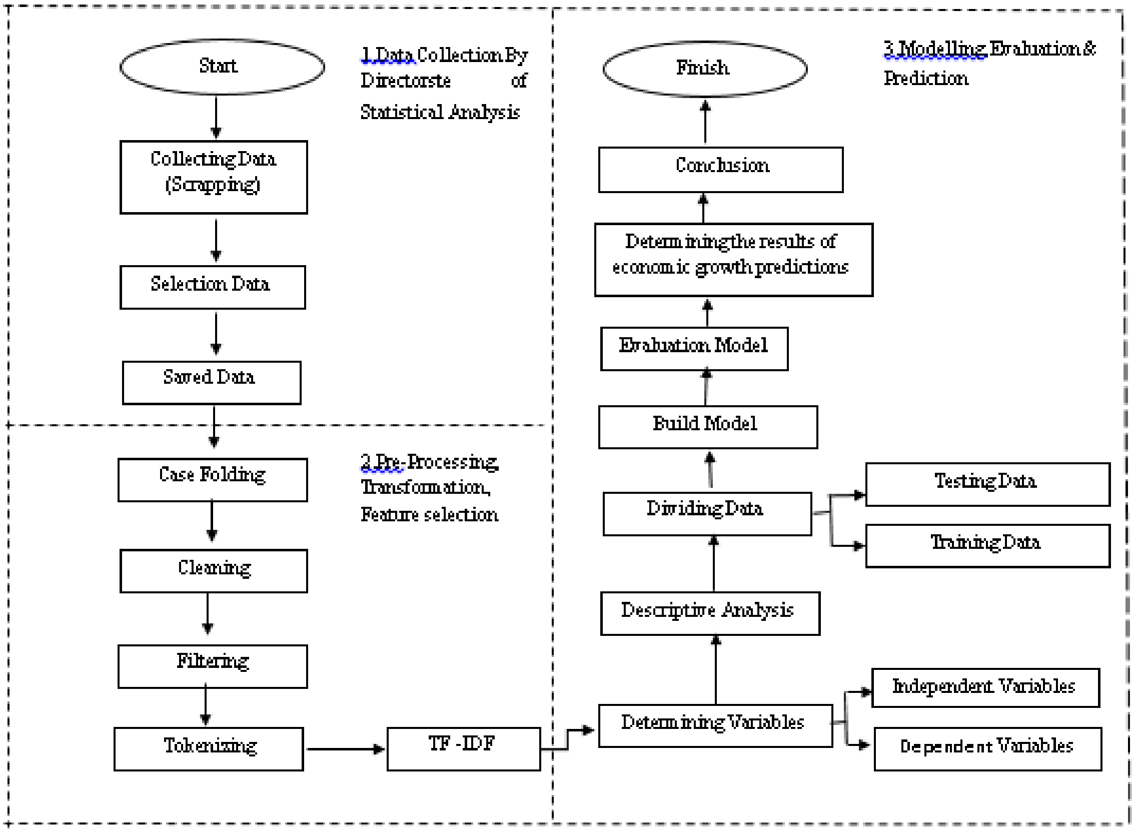

Figure 2. Stages of Analysis

Based on Figure 2, the stages of analysis in this study were carried out in several steps as follows:

1. Scrapping data on online news portals and grouped according to keywords for each category of business sector sub-sector.
2. It is pre-processing data or cleaning up unstructured data.
3. Doing term document or word weighting.
4. Labelling news that is up or down for variable Y, which the Directorate of Balance of BPS carries out
5. Perform descriptive analysis
6. Divide the data into two, with 80% for training data and 20% for randomly testing data.
7. Modelling using training data for the first quarter of 2021.
8. Make predictions using the training model that has been made with the data testing is the data for the second quarter of 2021.
9. Evaluate the model on the prediction results of the data obtained using the confusion matrix

RESULT

Statistic Descriptiv

This study makes predictions for forecasting the value of GDP in each category of GDP in the next quarter. The GDP category with the highest positive and negative growth values in the first quarter of 2021 is shown in Figure 3.

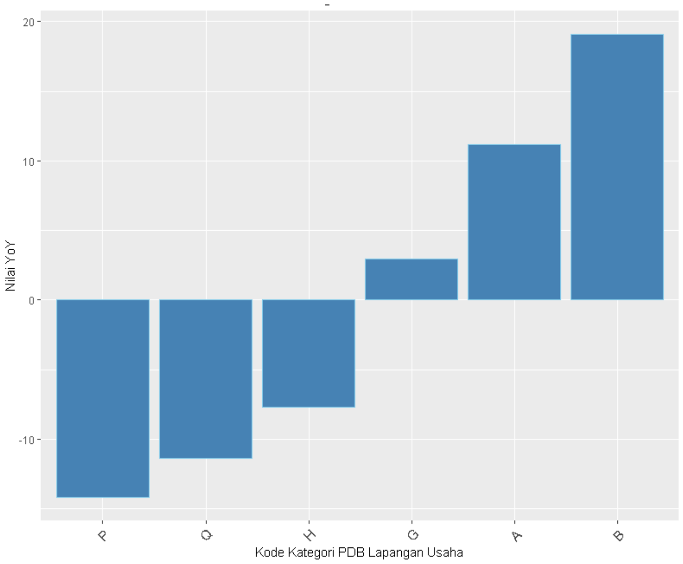

Figure 3 QoQ Value Category GDP of Quarter I

Based on the plot in Figure 3, category B, category A, and category G are the three categories that experienced positive growth. In contrast, Category P, category Q, and category H were three categories that experienced negative growth. so that to find out the state of economic growth in the second quarter of 2021 as an early warning, predictions of economic growth for each category are made using random forest

Prediction Results, Accuracy, Sensitivity, and Specificity

The classification modelling uses news data in the first quarter of 2021 in each category of GDP, with the dependent variable resulting from labeling, namely increasing or decreasing based on keywords validated by the BPS Balance Sheet Directorate team. The data modelling process for data classification analysis is divided into training data and testing data with a ratio of 80:20, which is done randomly, using cross-validation 10.

Table 8 Comparison of Classification Method Accuracy Results

MODEL	ACCURACY
Logistic Regression	95.26
Naive bayes	93.20
Random Forest	96.51
XGBoost	96.44
Support Vector Machine	96.28
SGD	96.06
Bernouli	93.84
MLP	96.23

Based on Table 8, the comparison of the classification models above, it is obtained that the Random Forest model has the highest accuracy. The random forest can reduce the variance of the final model, which for low variance means low overfitting. The random forest method performs well on datasets that do not balance between classes and can handle large amounts of data with higher variable dimensions. Still, in random forest modelling, one must pay attention that tree predictions must not be correlated, so in modelling economic growth predictions for each category of business field, a random forest model is used with a division of 80:20. The parameter value in the form of n estimator = 1000, i.e., the number of trees you want to make, before voting for the highest value of each prediction, a higher value of the number of trees gives a better classification performance. The results of the random forest classification can be seen in the confusion matrix in Table 9.

Table 9 Calculation Precision, Recall, F1 score Random Forest

	Precision $\frac{TN}{(TN + FP)}$	Recall $\frac{TP}{(TP + FN)}$	F1 Score $2\ \frac{Precision*recall}{(precision+recall)}$
Increase	0.97	0.98	0.98
Down	0.97	0.97	0.97
Rata-Rata	0.97	0.97	0.97

Based on Table 9, the accuracy value is 96.51%. Precision, i.e., 0.97 or 97%, indicates the number of correct or relevant predictions of all predictions based on the positive class. The recall value of 0.97 or 97% shows the number of negative misclassifications. The fewer negative misclassifications have been given, the higher the recall value and the average f1 score is also 0.97 or 97%, which indicates the weighted average accuracy measurement of precision and recall.

The results of forecasting GDP growth for the second quarter of 2021 can be seen in Table 10, which uses training data for the first quarter of 2021.

Table 10 Prediction of Data for The Second Quarter of 2021 Based on QoQ

GDP Category	**Actual Value**	**Predicted Value**
A. Agriculture, Forestry, and Fisheries	Increase	Increase
B. Mining and Quarry	Increase	Increase
C. Processing Industry	Increase	Increase
D. Electricity and Gas Procurement	Down	Down
E. Water Supply, Waste Management, Waste and Recycling	Increase	Increase
F. Construction	Down	Increase
G. Wholesale and Retail Trade; Car and Motorcycle Repair	Increase	Increase
H. Transportation and Warehousing	Increase	Down
I. Provision of Accommodation and Drinks	Increase	Increase
J. Information and Communication	Increase	Increase
K. Financial Services and Insurance	Increase	Increase
M N. Company Services	Increase	Down
P. Education Services	Increase	Increase
Q. Health Services and Social Activities	Increase	Increase

Based on Table 10, the predicted results of economic growth in each category of GDP, three categories were wrongly predicted, namely the Construction Category, the category of transportation and warehousing, and the category of company services. Classification errors occur because many new phenomena occurred in the second quarter of 2021 that did not exist in the previous quarter in the training data. Can add training data to the next modelling to draw general and representative conclusions. Based on Table 10, economic growth predictions in the second quarter of 2021 are predicted to increase.

DISCUSSIONS

The average precision value in this model is 0.97 or 97%, which shows that the entire class that should predict to go up and correctly predicted to go up is 97%, and 97% of the entire class that should be down correctly predicted to go down is 97%. The average recall value is 0.97 or 97%, which indicates the number of negative misclassifications given. The f1 score is also 0.97 or 97%, which indicates a weighted average accuracy measurement of precision and recall. The Random Forest model provides an accuracy of 96.51%, meaning that the model is very good at classifying data, so we can say that the random forest model has a good performance in predicting economic growth for the next quarter.

CONCLUSION

The results showed that the prediction of economic growth using online news training data in the 1st quarter of 2021 in each GDP category using the Random forest and cross-validation methods obtained an accuracy value of 96.51%, precision of 97%, and recall of 97%. The Random forest method is very efficient for classifying news in each GDP category. The results of this classification can predict economic growth in the next quarter, namely the second quarter of 2021, with predictions that the results will increase or experience positive growth. Suggestions for further research can be to predict economic growth based on the GDP category to get a specific value or quantitatively and predict using a Year On Year (YoY) comparison.

.

REFERENCES

Barua A, Sharif O, Hoque MM, Barua A, Sharif O, Hoque MM. 2021. Multi-class Sports News Categorization Categorization using Machine Learning Techniques : Resource Creation and Evaluation T. Procedia Computer Science. 193:112–121. doi:10.1016/j.procs.2021.11.002.

Brookes, B. C. (1972). The Shannon model of IR systems. Journal of Documentation, 28, 160–162.

Bramer, M. (2007). Principles of Data Mining. In M. Bramer (Ed.), Springer Science+Business Mediaspringer.com (Issue January 2007). Springer Science+Business Media springer.com. https://doi.org/10.1007/978-1-84628-766-4

Cover, T. M., & Thomas, J. A. (1991). Elements of information theory. New York: John Wileyand Sons Inc

Dhar P, Abedin MZ. 2021. Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. I.J Informatics Enginering Electronic. Busi. 13(1):15–24. doi:10.5815/ijieeb.2021.01.02.

Hastie et al. 2008. The Elements of Statistical Learning. Elem Stat Learn. 26(4):505–516.

Leo Breiman. 2001. Random Forests. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 12343 LNCS:503–515. doi:10.1007/978-3-030-62008-0_35.

Manning C.D, Raghavan P, Scutze, H. 2009. Introduction to Information Retrieval. Ed ke-1. Cambridge: Cambridge University Press

Lakshmanaprabu, S., Shankar, K., Ilayaraja, M., Nasir, A.W., Vijayakumar, V., Chilamkurti, N., 2019. Random forest for big data classification in the internet of things using optimal features. Int. J. Mach. Learn. Cybern. 10 (10), 2609–2618.

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition Elsevier Ltd, 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2017). Machine learning: Algorithms and applications. In E. B. M. B. Mohssen Mohammed, Muhammad Badruddin Khan (Ed.), Machine Learning: Algorithms and Applications. CRC Press Taylor & Francis Group. https://doi.org/10.1201/9781315371658.

Parida, U., Nayak, M., Nayak, A.K., 2021. News text categorization using random forest and naïve bayes. In: 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON), IEEE, pp. 1–4.

Pathak MA. 2014. Beginning Data Science with R. Springer.

Sarkar D. 2019. Text Analytics with Python. India :Apress.

U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev, A. Alizade, Empirical study of online news classification using machine learning approaches, in: IEEE Int. Conf. on Application of Information and Communication Technologies, IEEE, 2018, pp. 1–6.

Wong, S. K. M., & Yao, Y. Y. (1992). An information-theoretic measure of term specificity. Journal of the American Society for Information Science, 43(1), 54–61

Wongso R, Luwinda FA, Trisnajaya BC, Rusli O, Wongso R, Luwinda FA, Trisnajaya BC, Rusli O. 2017. News Article Text Classification in Indonesian Language. Procedia Computer Science. 116:137–143. doi:10.1016/j.procs.2017.10.039.

Zhu M. 2008. Kernels and ensembles: Perspectives on statistical learning. Am Stat. 62(2):97–109. doi:10.1198/000313008X306367