BAREKENG: Journal of Mathematics and Its Applications
June 2023 Volume 17 Issue 2 Page 0847–0854
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss2pp0847-0854

CABLE NEWS NETWORK (CNN) ARTICLES CLASSIFICATION USING RANDOM FOREST ALGORITHM WITH HYPERPARAMETER OPTIMIZATION

Krisna Sidiq[1], Dewi Retno Sari Saputro[2*]

[1,2]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret*
Ir. Sutami Street No. 36, Surakarta, 57126, Indonesia

*Corresponding author's e-mail: * **dewiretnoss@staff.uns.ac.id***

Article History:

Received: 20th December 2022
Revised: 15th April 2023
Accepted: 19th April 2023

Keywords:

Classification;
Hyperparameter;
Random forest.

ABSTRACT

The growth of news articles on the internet occurs in a short period with large amounts, so necessary to be grouped into several categories for easy access. There is a method for grouping news articles, namely classification. One of the classification methods is a random forest which is built on decision tree. This research discusses the application of random forest as a method of classifying news articles into six categories; these are business, entertainment, health, politics, sport, and news. The data used is Cable News Network (CNN) articles from 2011 to 2022. The data is in the form of text and has large amounts, so good handling is needed to avoid overfitting and underfitting. Random forest is proper to apply to the data because the algorithm works very well on large amounts of data. However, random forest has a difficult interpretation if the combination of parameters is not appropriate in the data processing. Therefore, hyperparameter optimization is needed to discover the best combination of parameters in the random forest. This research uses the search cross-validation (SearchCV) method to optimize hyperparameters in the random forest by testing the combinations one by one and validating those. Then we obtain the classification of news articles into six categories with an accuracy value of 0.81 on training and 0.76 on testing.

How to cite this article:

K. Sidiq and D. R. Sari Saputro., "CABLE NEWS NETWORK (CNN) ARTICLES CLASSIFICATION USING RANDOM FOREST ALGORITHM WITH HYPERPARAMETER OPTIMIZATION," *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 0847-0854, June, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

The growth of news articles on the internet occurred in a short period with large numbers [1]. A large number of news articles need to be grouped into several categories for easy access to news articles [2]. News articles can be grouped manually, but it takes a long time. There is a method for grouping news articles, namely classification. Classification is the process of identifying and grouping objects or ideas into predetermined categories. Classification of news articles is included in text mining because it requires preparatory steps to convert unstructured data into structured information [3].

Research related to the classification of news articles has been carried out using various algorithms and methods. Dion and Kartika (2015) researched to classify news articles using the Naïve Bayes algorithm and support vector machine (SVM), showing good levels of accuracy, which are 82,2% and 88,1% [4]. Fanny (2018) conducted another research using the k-nearest neighbor (KNN) algorithm with the type of correlation similarity obtained good results, namely 86,11% on stemmer evaluation [5]. However, all three algorithms have shortcomings in classifying large amounts of news article data. Naïve Bayes has very strong independence between features, thereby reducing the correlation between features that play an important role in determining categories. SVM has a weakness in processing large amounts of data, it causes frequent overfitting or underfitting. Meanwhile, KNN cannot do learning well if the type of attribute used is not in accordance with the dataset.

The dataset used in this research is Cable News Network (CNN) news articles from 2011 to 2022. CNN is a multi-national cable news channel that provides 24-hour news coverage. The dataset has a large size and strong correlation between its features, so it is not appropriate to use algorithms such as Naïve Bayes, SVM, and KNN for classification. Therefore, in this research, the random forest algorithm is used, which is an ensemble and multi-class classification algorithm. Random forest is an algorithm that is built from a combination of various decision trees [6]. The advantage of random forest is that it can work very well on large amounts of data. In addition, random forests can estimate features that are important in the classification process and provide experimental methods to detect correlations between features [7].

Random forest algorithm is prone to overfitting or underfitting if the combination of parameters used is not appropriate for processing the data. Therefore, in this research, hyperparameter optimization was carried out to find the best combination of parameters in the random forest algorithm. This research uses the search cross-validation (SearchCV) method, which is a method of selecting a combination of parameters and random forest algorithms by testing the combinations one by one and validating them [8].

2. RESEARCH METHODS

This research was conducted in four main stages, which are data acquisition, data preprocessing, classification, and evaluation.

2.1 Data Acquisition

This research uses data from Cable News Network (CNN) news articles from 2011 to 2022 [9]. The number of rows in the dataset before preprocessing is 37.949 rows of data. The dataset has 11 columns, these are Index, Author, Date published, Category, Section, Url, Headline, Description, Keywords, Second headline, and Article text. Table 1 shows the top five rows of data used in this research before data preprocessing.

Table 1. Cable news network (CNN) articles dataset

Index	Author	Date published	Category	Section	Url	Headline	Description	Keywords	Second headline	Article text
0	Jacopo Prisco,	2021-07-15	news	world	https://www.cnn.com/2021/07/14/...	shortage truckers tusimple think solution driv...	e commerce boom exacerbate global truck driver...	world, There's a shortage of truckers, but TuS...	shortage truckers tusimple think solution driv...	cnn right shortage truck drivers us worldwide ...

Index	Author	Date published	Category	Section	Url	Headline	Description	Keywords	Second headline	Article text
1	Stephanie Bailey,	2021-05-12	news	world	https://www.cnn.com/2021/05/12/...	bioservo robotic ironhand could protect factor...	work factory mean task could lead chronic inju...	world, Bioservo's robotic 'Ironhand' could pro...	robotic ironhand could protect factory workers...	cnn work factory warehouse mean task repetitio..
2	words Stephanie Bailey, video Zahra Jamshed	2021-06-16	news	asiaS	https://www.cnn.com/2021/06/15/...	swarm robots get smarter work cnn	hong kong warehouse swarm autonomous robots wo...	asia, This swarm of robots gets smarter the mo...	swarm robots get smarter work	cnn hong kong warehouse swarm autonomous robot...
3	Kathryn Vasel	2022-03-18	business	success	https://www.cnn.com/2022/03/18/...	two years later remote work change millions ca...	look pandemic reshape people career ways never...	success, Two years later, remote work has chan...	two years later remote work change millions ca...	pandemic thrust work world new reality march t...
4	Paul R la Monica,	2022-03-19	business	investing	https://www.cnn.com/2022/03/19/...	march volatile stock cnn	march madness college basketball fan phrase al...	investing, Why March is so volatile for stocks...	march volatile stock	new york cnn business march madness college ba...

***Data source**: Kaggle*

Based on **Table 1**, the variables used in this research are Index, Category, and Article Text. In this research, the Article Text variable in the form of text data is categorized based on the classes available in the Category variable. In this research, news articles were grouped into six categories, namely news, sport, politics, business, health, and entertainment, because the vr, travel, and style categories have a very small number of rows of data. The amount of data that is not balanced can make the results of the algorithm's accuracy skew towards the majority object **[10]**. Therefore, the category with the smallest number of rows needs to be deleted and balanced by the sampling method.

2.2 Data Preprocessing

At the data preprocessing stage, various data handling processes are carried out. The handling process aims to ensure good data quality before being used during data analysis. Several things need to be ensured, namely data accuracy, completeness, consistency, timeliness, reliability, and being able to be interpreted **[11]**. In this research, data preprocessing was carried out on text data, including case folding, converting numbers to words, removing special characters, removing stop words, lemmatization, stemming, split data training and testing, vectorization with TF-IDF, and data balancing.

2.3 Classification

Random forest algorithm is built from a collection of several decision trees. A decision tree is a tree-shaped structure that has several parts, namely the root node, which is used to collect data, the inner node which contains data questions, and the leaf node which is used to solve problems and make decisions **[12]**. Prediction results from the random forest are obtained through the highest results from each decision tree (voting for classification), as shown in **Figure 1**. For random forest consisting of N trees, **Equation (1)** is used to predict the class l label of the case y through voting **[7]**.

$$l(y) = argmax_c\left(\sum_{n=1}^{N} I_{h_n(y)=c}\right) \quad (1)$$

Where I is the indicator function and h_n is the n-th tree of the random forest.

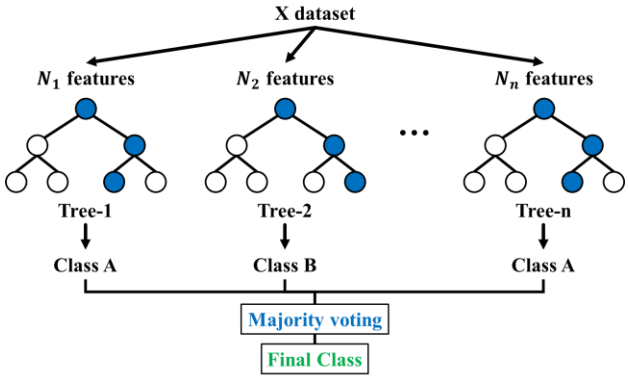

Figure 1. Random forest architecture

Random forest algorithm is more accurate in estimating the error rate than the decision tree. In particular, the error rate has been proven mathematically to always converge as the number of trees increases. To be able to produce accurate and stable predictions, random forest works by applying the bagging method (bootstrap aggregation). The bagging method is a collection of several meta-algorithms that aim to improve the accuracy of machine learning algorithms [13]. Random forest can work efficiently when applied to large-scale datasets with high accuracy and easy-to-understand results. However, random forest requires setting the right combination of parameters in the dataset to avoid cases of overfitting or underfitting. Ramadhan *et al.*, in their research, using SearchCV method to find the best combination of parameters in random forest model [8]. In addition, Adnan *et al.* also use SearchCV to improve the performance of the classification model that has been created. SearchCV is a method of selecting a combination of parameters and models by testing the combinations one by one and validating each combination to produce the best model performance [14].

2.4 Evaluation

Evaluation is the stage of measuring the performance of a model that has been made so that it can be considered in choosing the best model. In this research, the evaluation metrics used are the value of accuracy and confusion matrix. The accuracy value is obtained by dividing the number of correct predictions by the total number of predictions. However, accuracy can be misleading if there is a large class imbalance, so an additional evaluation metric is used, namely the confusion matrix. The confusion matrix displays and compares the actual value with the predicted value in the classification case. The confusion matrix displays and compares the actual value with the predicted value in a classification case. In the research conducted by Hossin and Sulaiman, the confusion matrix helps in knowing the comparison between prediction errors and prediction accuracy in each classification class in detail [15].

3. RESULTS AND DISCUSSION

In this research, the data is divided into two parts, namely training data and testing data with a proportion of 85:15, then vectorization is carried out. Vectorization is the process of converting text data into numeric data. The vectorization method used is TF-IDF (Term Frequency-Inverse Document Frequency) [16]. Before training the data, it is necessary to first check the proportion of each category/class to avoid data imbalance.

Table 2. Comparison of data before and after balancing

Category	Index	Preliminary data	Data after balancing
News	3	15358	351
Sport	5	13210	351
Politics	4	2092	351
Business	0	726	351
Health	2	473	351
Entertainment	1	351	351

From the search results shown in **Table 2**, the data is not balanced because it is dominated by the "news" and "sport" categories. Data imbalance can cause overfitting or underfitting in the classification process so that low accuracy is obtained. Overfitting occurs when the model learns the training data too well, while underfitting does not study the training data well. Therefore, it is necessary to balance the data. The library used in this study is "Imblearn". "Imblearn" is a method for balancing data in each class so that it has the same amount, with the random undersampling method, which reduces the data in the majority category/class **[17]**. Random forest algorithm used for classification is set to random_state=5. **Figure 2** shows metrics for evaluating the classification results using the random forest algorithm without hyperparameter optimization. Numbers 0, 1, 2, 3, 4, and 5 in the classification report indicate the index of each category, as shown in **Table 2**.

```
Accuracy score on train:  1.0
Accuracy score on test:  0.7903225806451613
Classification report:
              precision    recall  f1-score   support

           0       0.78      0.76      0.77        62
           1       0.69      0.84      0.76        62
           2       0.77      0.81      0.79        62
           3       0.73      0.56      0.64        62
           4       0.85      0.84      0.85        62
           5       0.92      0.94      0.93        62

    accuracy                           0.79       372
   macro avg       0.79      0.79      0.79       372
weighted avg       0.79      0.79      0.79       372
```

Figure 2. Metrics for evaluating random forest algorithm classification results without hyperparameter optimization

In **Figure 2**, it can be seen that the accuracy value of the training data is perfect, which is 1, while the accuracy value of the testing data is 0,79. It can be concluded that the model of random forest algorithm that has been made is overfitting. Random forest is indeed suitable to be applied to large amounts of data. However, without proper handling often results in overfitting or underfitting. Therefore, it is necessary to handle the random forest algorithm, one of which is by optimizing hyperparameters **[18]**. Hyperparameter optimization is a technique to find a combination of parameters with the random forest algorithm so that the best classification results are obtained **[19]**. **Table 3** shows the combination of parameters used in the random forest algorithm training.

Table 3. Combination of random forest default parameters

Parameter	Value
max_depth	5
min_samples_leaf	2
n_estimators	4

In **Table 3**, it can be seen that the random forest algorithm by default regulates the combination of parameters used.

In this research, a search for the best combination of parameters in the random forest algorithm will be carried out. The method used to find the combination of parameters is search cross-validation (SearchCV) **[8]**. The parameters used are n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, and bootstrap. The best combination of parameters is then fitted to the random forest algorithm. **Table 4** shows the comparison of various combinations of parameters and the level of accuracy of the classification results.

Table 4. Comparison of accuracy levels on various combinations of parameters

Parameter	Combination 1	Combination 2	Combination 3	Combination 4	Combination 5
n_estimators	576	354	919	253	456
max_features	'sqrt'	'auto'	'sqrt'	'auto'	'sqrt'

Parameter	Combination 1	Combination 2	Combination 3	Combination 4	Combination 5
max_depth	5	5	5	5	5
min_samples_split	2	4	4	2	2
min_samples_leaf	3	5	5	4	2
bootstrap	True	True	True	False	True
Accuracy on train	0,81	0,8	0,81	0,8	0,81
Accuracy on test	0,74	0,73	0,75	0,72	0,76

Each value set for each parameter has a significant impact on the level of accuracy. The number of values for each parameter will produce good accuracy, but it takes a long time, and vice versa. In **Table 4**, it can be seen that Combination 5 has the best accuracy, and the distance between the accuracy of the training data and the testing data is neither too far nor too close. **Figure 3** shows the evaluation metrics for classification results using Combination 5.

```
Accuracy score on train:  0.8133903133903134
Accuracy score on test:  0.7580645161290323
Classification report:
              precision    recall  f1-score   support

           0       0.76      0.55      0.64        62
           1       0.65      0.84      0.73        62
           2       0.73      0.87      0.79        62
           3       0.80      0.52      0.63        62
           4       0.78      0.82      0.80        62
           5       0.87      0.95      0.91        62

    accuracy                           0.76       372
   macro avg       0.76      0.76      0.75       372
weighted avg       0.76      0.76      0.75       372
```

Figure 3. Metrics for evaluating random forest algorithm classification results with hyperparameter optimization

Figure 3 shows that the classification of news articles into six categories has been successfully carried out. This can be seen from the accuracy value of 0.81 in training and 0.76 in testing so that the model created is protected from overfitting and underfitting. This value also shows that the model is able to classify well, namely 76% of new data can be classified correctly. In addition, **Figure 4** shows the confusion matrix for each classification class.

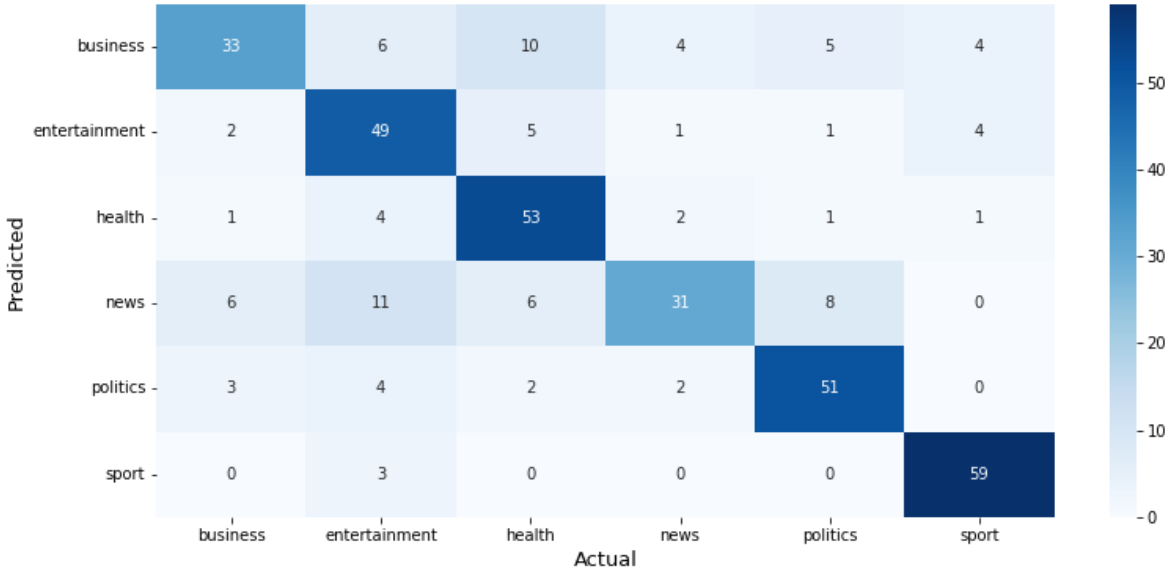

Figure 4. Confusion matrix

Figure 4 shows the amount of accuracy and prediction error of the data in each classification class. The sport category is the class with the highest number of correctly predicted data, namely 59 data, while the news is the class with the smallest number of correctly predicted data, namely 31 data. Overall, the model made is able to classify with fairly good accuracy; it can be seen from the diagonal which has a dark color. Thus, the random forest algorithm with hyperparameter optimization can classify news articles into six categories.

4. CONCLUSIONS

News articles need to be grouped into several categories for easy access for readers. One of the algorithms that can be used for classification problems is random forest. In this research, the random forest algorithm worked well in classifying CNN news articles into six categories, namely news, sport, politics, business, health, and entertainment. Hyperparameter optimization on the random forest algorithm has a significant impact on the classification results where the method used is randomized search cross-validation. This method can find the best combination of parameters in the random forest algorithm at random. However, keep in mind that determining the combination of parameters is much better if the value set for each parameter is varied and numerous, but it takes a long time. In addition, hyperparameter optimization aims to avoid overfitting, underfitting, and inappropriate processing of training data. This research obtained classification results with an accuracy value of 0,81 on training and 0,76 accuracy on testing. Thus, this research can be used as a reference in classifying news articles for easy access.

ACKNOWLEDGMENT

The authors would like to thank Mathematical Soft Computing Research Group for their moral support and incisive comments to improve this article. The research described in this paper is supported by a fundamental grant from the Research and Community Service Institute of Universitas Sebelas Maret through a letter of agreement for the implementation of the research implementation of the non-APBN Fund No. 254/UN27.22/PT.01.03/2022.

REFERENCES

[1] House of Lords, "Growing up with the internet," *Parliament.uk*, no. March, 2017, [Online]. Available: https://www.publications.parliament.uk/pa/ld201617/ldselect/ldcomuni/130/13002.htm.

[2] O. de Clercq, L. de Bruyne, and V. Hoste, "News topic classification as a first step towards diverse news recommendation," *Comput. Linguist. Netherlands J.*, vol. 10, pp. 37–55, 2020.

[3] G. Kaur and K. Bajaj, "News Classification and Its Techniques: A Review," *IOSR J. Comput. Eng.*, vol. 18, no. 1, pp. 22–26, 2016, doi: 10.9790/0661-18132226.

[4] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. Sains dan Seni ITS*, vol. 4, no. 2, pp. 248–253, 2015.

[5] Fanny, Y. Muliono, and F. Tanzil, "A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification," *J. Inform. J. Pengemb. IT*, vol. 3, no. 2, pp. 157–160, 2018, doi: 10.30591/jpit.v3i2.828.

[6] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[7] D. Liparas, Y. HaCohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "News Articles Classification Using Random Forests and Weighted Multimodal Features BT - Multidisciplinary Information Retrieval," 2014, pp. 63–75.

[8] M. Ramadhan, I. Sitanggang, F. NASUTION, and A. GHIFARI, "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency," in *DEStech Transactions on Computer Science and Engineering*, Oct. 2017, pp. 625–629, doi: 10.12783/dtcse/cece2017/14611.

[9] H. Unger, "CNN News Articles from 2011 to 2022," *Kaggle*, 2022. https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning.

[10] R. Siringoringo, "Klasifikasi Data tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[11] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.

[12] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[13] N. Altman and M. Krzywinski, "Ensemble methods: Bagging and random forests," *Nat. Methods*, vol. 14, no. 10, pp. 933–934, 2017, doi: 10.1038/nmeth.4438.

[14] M. Adnan, A. A. S. Alarood, M. I. Uddin, and I. ur Rehman, "Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models," *PeerJ Comput. Sci.*, vol. 8, no. Ml, pp. 1–29, 2022, doi: 10.7717/PEERJ-CS.803.

[15] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[16] A. Hakim, A. Erwin, I. E. Kho, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," Jan. 2015, doi: 10.1109/ICITEED.2014.7007894.

[17] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, 2017.

[18] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1301, 2019, doi: https://doi.org/10.1002/widm.1301.

[19] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, doi: 10.3390/informatics8040079.