

Experimental of information gain and AdaBoost feature for machine learning classifier in media social data

Jasmir Jasmir¹, Dodo Zaenal Abidin², Fachruddin Fachruddin³, Willy Riyadi¹

¹Department Computer Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

²Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

³Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

Article Info

Article history:

Received Mar 14, 2024

Revised Jul 29, 2024

Accepted Aug 5, 2024

Keywords:

AdaBoost
Information gain
Machine learning
Social media
Text classification

ABSTRACT

In this research, we use several machine learning methods and feature selection to process social media data, namely restaurant reviews. The selection feature used is a combination of information gain (IG) and adaptive boosting (AdaBoost) which is used to see its effect on the classification performance evaluation value of machine learning methods such as Naïve Bayes (NB), K-nearest neighbor (KNN), and random forest (RF) which is the aim of this research. NB is very simple and efficient and very sensitive to feature selection. Meanwhile, KNN is known for its weaknesses such as biased k values, overly complex computation, memory limitations, and ignoring irrelevant attributes. Then RF has weaknesses, including that the evaluation value can change significantly with only small data changes. In text classification, feature selection can improve the scalability, efficiency and accuracy of text classification. Based on tests that have been carried out on several machine learning methods and a combination of the two selection features, it was found that the best classifier is the RF algorithm. RF produces a significant increase in value after using the IG and AdaBoost features. Increased accuracy by 10%, precision by 12.43%, recall by 8.14% and F1-score by 10.37%. RF also produces even accuracy, precision, recall, and F1-score values after using IG and AdaBoost with an accuracy value of 84.5%; precision of 85.58%; recall was 86.36%; and F1-score was 85.97%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Jasmir Jasmir

Department Computer Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa

Jendral Sudirman Street, Tehok, South Jambi, Jambi, Indonesia

Email: ijay_jasmir@yahoo.com

1. INTRODUCTION

Currently, there is a proliferation of computerized texts, flooding our digital landscape. Each day witnesses the emergence of numerous new web pages, alongside a continuous stream of news articles, magazine pieces, and scholarly writings, particularly on social media platforms. This surge results in an abundance of textual content available in digital form [1], [2]. With digital texts being widely accessible and the demand for flexible access continually growing, the task of text classification has become indispensable [3]. However, one of the primary challenges in this domain lies in the vast dimensionality of the feature space [4]. Many of these features prove irrelevant or detrimental to classification accuracy, necessitating the identification and incorporation of more relevant features to enhance performance [5].

Due to the large amount of unstructured information available on the Web, gathering and compiling information is a challenging task, requiring the use of automated methods to help researchers collect and

analyze sentiment-related data [6]. The object of sentiment analysis can be speech, text, and images. Here we use a restaurant review dataset which is usually presented in text form, so sentiment analysis in most papers focuses on text-based sentiment analysis [7].

Several previous studies related to sentiment analysis, as discussed by Muktafin and Kusriani [8] discussed sentiment analysis of public service customer satisfaction using K-nearest neighbor (KNN), with the TF-IDF feature and produced an accuracy of 74%. Then Santoso *et al.* [9] discussed sentiment analysis of hoax news using Naïve Bayes (NB) and produced an accuracy of 77%. Meanwhile, Khalid *et al.* [10] discussed sentiment analysis for the spread of COVID-19 using deep learning, namely bidirectional long-short term memory, and produced an accuracy of 74.92%. Looking at the results of the classification performance evaluation values above, there is still an opportunity to increase the classification performance evaluation values by carrying out several experiments using machine learning methods combined with several features to increase the classification performance evaluation values.

One of the common aspects in sentiment classification approaches is feature selection. The process of selecting features can enhance both the efficiency and the efficacy of the classifier [11]. This can be achieved by diminishing the volume of analyzed data and pinpointing pertinent features that should be taken into account during the learning process. One of the superior features is Information gain (IG) [12]. IG evaluates how much a word's presence or absence aids in making precise classification decisions across all categories. It serves as an effective filtering method in text classification [13], [14].

Apart from the IG feature, there is another feature such as adaptive boosting (AdaBoost). AdaBoost is an algorithm whose basic concept is to select and combine a group of weak classifiers to form a strong classification [15]. The AdaBoost algorithm is specifically designed for classification purposes, where the learning is aimed at improving the accuracy of any weak learning algorithm [11]. AdaBoost is used generally to improve the accuracy of weak learning in partially supervised learning classification tasks [16]. The machine learning methods we use are NB, KNN, and random forest (RF). However, NB still has shortcomings, namely when dealing with complex dimensions, it will result in a low level of classification accuracy and produce biased classification results [17]. Meanwhile, KNN has disadvantages, including being very dependent on feature scaling [18], [19]. RF has a weakness, namely that to achieve predictions with a high level of accuracy, more computing resources are needed. The greater the need for resources, the longer it takes to produce predictions [20], [21].

Therefore, referring to the problems above, we conducted research as well as the contribution of this research, namely to increase the accuracy of several machine learning methods, namely NB, KNN, and RF by using the IG and AdaBoost features as a technique to increase the classification performance evaluation value of machine learning on restaurant review datasets.

2. METHOD

This research has a big emphasis on the experimental stage to determine the effect of the IG feature and the AdaBoost feature in increasing the evaluation value of machine learning classification performance. These experiments involve a thorough analysis of several components, including the use of feature selection techniques, the allocation of data sets for training and testing, and the design and setup of machine learning methods and associated variables. To ensure optimal results for this research, a structured series of important steps was designed to develop an appropriate model and avoid deviation from the intended goals. The classification process is outlined in Figure 1.

- Initially, data collection required the use of publicly available datasets from Kaggle.
- Next, the text undergoes preprocessing, an important step in preparing it for training and testing, which involves tokenization, stop word removal, and stemming.
- After preprocessing, the research moves on to conducting training and testing, which involves two approaches: one without utilizing features and the other using features. In the feature-based approach, a combination of IG and AdaBoost features is applied along with three classifiers. The analysis considers parameters such as accuracy, precision, recall, and F1-score. Testing was carried out using the results of IG and AdaBoost.
- After implementing all the methodologies, the next step is to compare the results of the training and testing processes of the two approaches. Each classifier integrates a mix of IG and AdaBoost features.
- The process ends with an evaluation of the training and testing procedures, as well as an analysis of the resulting classification performance.

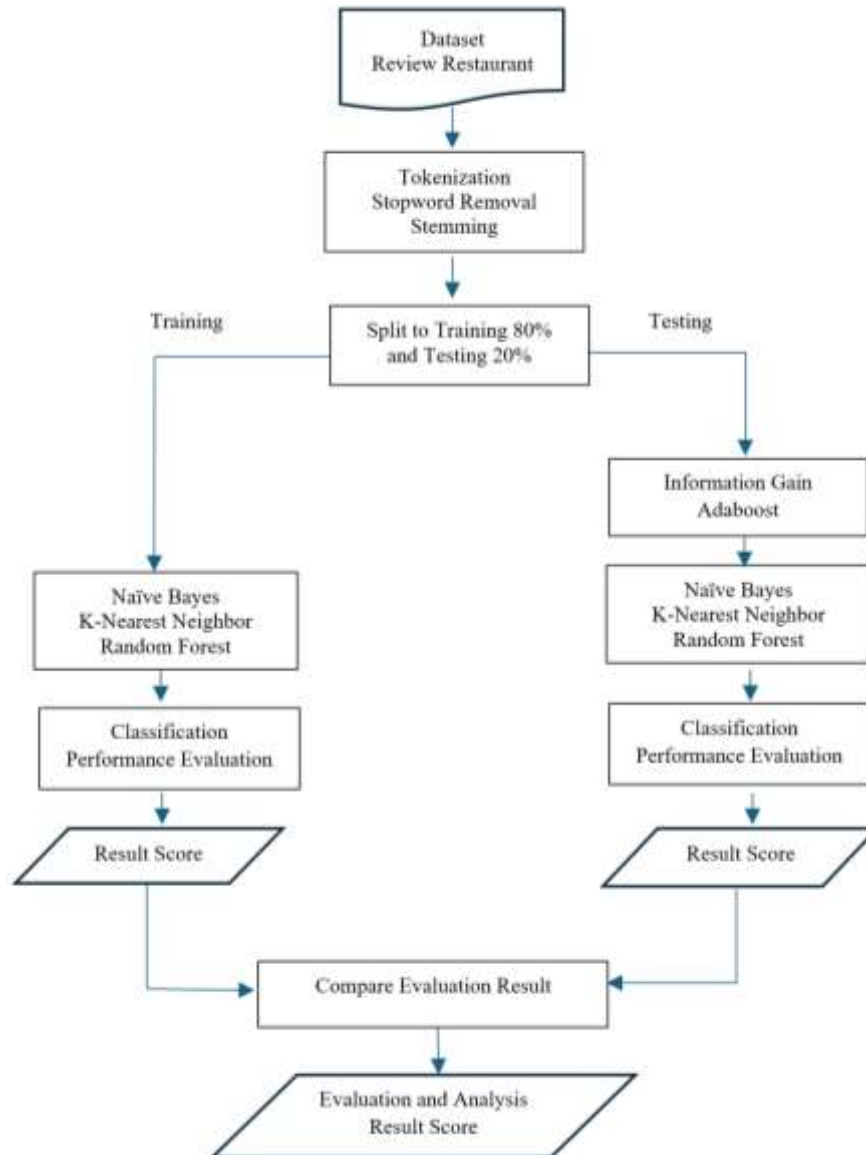


Figure 1. Research framework

2.1. Dataset

The selection of a dataset is influenced by the data to be processed, including searching for existing data, and obtaining additional data as needed. We use open dataset from kaggle. Overall, we have discovered some interesting insights that could prove useful for restaurant owners. Next, integration of the collected data occurs in data sets. In this study, a restaurant review dataset was selected. There are 200 restaurant review data consisting of 100 positive review data and 100 negative review data [22].

2.2. Preprocessing

Data selection was carried out. The data is cleaned and transformed into the desired form before making the model. The utilized dataset comprises solely 100 positive reviews and 100 negative reviews, serving as the training data. This dataset undergoes preprocessing through three distinct processes: Tokenization, Stopwords Removal, and Stemming.

2.3. Feature selection and boosting

IG stands as the most straightforward feature selection method, commonly employed for ranking attributes. Its widespread application extends to text categorization, microarray data analysis, and image data analysis [23], [24]. IG aids in mitigating noise stemming from irrelevant features by identifying those with

the most information relative to a specific class. The process of determining the optimal attribute involves initially calculating the entropy value, where entropy serves as a measure of uncertainty and can be utilized to deduce the concise distribution of features.

AdaBoost stands as one among several variations of the boosting algorithm [11]. AdaBoost, a form of ensemble learning frequently employed in boosting algorithms, can be integrated with other classifier algorithms to enhance overall classification performance. The intuitive idea is that combining diverse models proves beneficial. AdaBoost and its variations have found success across various domains, owing to their robust theoretical foundation, precise predictions, and straightforward implementation.

2.4. Learning model

2.4.1. KNN

KNN algorithm is an approach to classify objects by utilizing learning data that is in close proximity to the given object [25], [26]. KNN is a learning algorithm that does not involve a dedicated offline training phase [27]. All training documents are stored and computations are postponed until the prediction phase. In the case of each test document, KNN arranges the labelled examples from the training set based on proximity and utilizes the category of the highest-ranked neighbor to assign classes. The greater the proximity of neighbor within the same category, the more confident the prediction becomes [28].

2.4.2. Naïve Bayes

NB stands out as one of the most straightforward probabilistic classifier models [29]. This naïve assumption can offer a balanced compromise between performance and computational expenditure. Tang *et al.* [30] found that NB can also perform well when features are interdependent. In addition, Elhadad *et al.* [31] contend that the generative model produced is easily interpretable and explainable. Moreover, as a generative classifier, NB can be deemed suitable for smaller sample sizes owing to its inherent regularization, reducing the likelihood of overfitting when compared to discriminative classifiers [32]. Nevertheless, NB fails to capture interaction effects among features. As a result, it is anticipated to exhibit satisfactory performance in scenarios characterized by distinct individual signal words and straightforward connections between text features and their corresponding classes, such as in basic forms of promotional content detection [33].

2.4.3. Random forest

RF is an ensemble learning technique that constructs numerous random, independent decision trees (DTs). Each DT contributes a vote for the class of test examples, and the most prevalent class ultimately dictates the final prediction of the RF classifier. This process is referred to as bagging [34]. The greater the number of predictors, the more trees need to be generated to achieve optimal performance. Various techniques can be employed to incorporate randomness and enhance individual DTs, such as random feature selection and the random selection of subsets of the data. Despite their susceptibility to overfitting due to their high flexibility, these individual DTs can be improved through these randomization methods [35]. RF addresses this challenge by aggregating multiple DTs based on a randomly selected and diverse subset of variables.

3. RESULTS AND DISCUSSION

This research investigates the effects of the IG and AdaBoost features and their comparison with the NB, KNN, and RF methods in increasing the evaluation value of classification performance. While previous research has used KNN with TF-IDF features, the accuracy is still not very high. This research has explored the impact of the IG feature and the AdaBoost feature on increasing the classification performance evaluation value of the 3 methods above, but has not explicitly discussed its effect on the computational performance of the process. The results of this experiment are in line with the research objectives, showing an increase in classification performance evaluation values, namely accuracy, precision, recall, and F1-score. The data were divided into training data and testing data with a ratio of 80:20, and 10-fold cross-validation was used. Using the Rapidminer tool, the study was configured using the “select by weight” operator, with parameters set to “weight relationship = top k” and “k = 10”. Where the top 10 attributes and their respective weights will be generated as shown in Table 1.

The weights in Table 1 are the weights that have been generated by the select by weight operator. Because the result still has a value of 0, the only attribute whose weight is displayed in each document has a weight = 1. Among the 10 attributes above, only the dissapoint attribute has a weight = 1. The other nine attributes have a value below 1. Table 2 shows the attributes these attributes are in the document in vector form.

Table 1. Top features and their weights

Attribute	Weights
overpr	0.575
want	0.576
review	0.593
favorit	0.708
amaz	0.713
delici	0.713
good	0.767
definit	0.911
dissappoint	1

Table 2. Attributes in vector form

No	Number of documents	Disspoint	Class
1	12	2	Negative
2	17	2	Negative
3	28	2	Negative
4	96	2	Negative
5	23	1	Negative
6	64	1	Positive
7	76	1	Positive
8	149	1	Positive
9	64	1	Positive
10	76	1	Positive

We found that the IG features and AdaBoost features that were correlated with the NB, KNN, and RF methods tended to have a higher proportion of results than using machine learning methods without features. Table 3 presents the confusion matrix results from the NB method without using the IG feature and the AdaBoost feature and after using the IG feature and the AdaBoost feature. Before using this feature, you can see the results of the composition of the matrix values. If you look at these results, the false positive, and false negative values are still high. This can also happen because NB ability to process complex data is still not optimal, plus the composition of the attributes contained in the dataset cannot yet be used because most of the feature weights are still worth 0. Meanwhile, the results after using the features, the results are false positive and false negative has decreased compared to the previous false positive and false negative values. The most significant decrease was false positive, resulting in an increase in classification performance values. In this process, the influence of IG and AdaBoost can be proven, but the accuracy results are not yet significant. Even though they have improved, IG and AdaBoost cannot yet be categorized as suitable features for the NB method.

After carrying out NB testing with the IG and AdaBoost features, to see an overview of the differences in these results can be seen in Table 4 and Figure 2. From Table 4 and Figure 2 it can be seen that the results of the NB test without using the IG and AdaBoost features obtained values accuracy of 77.5%; precision 80.39%; recall of 76.63%; and F1-score of 78.46%. After using IG and AdaBoost, the accuracy value increased by 80.5%; precision to 84.15%; recall became 78.70%; and F1-score became 81.34%. This means that the NB algorithm experienced an increase in accuracy of 3%, precision of 3.76%, recall of 2.06%, and F1-score of 2.87%. From the confusion matrix above, each classification performance evaluation value is obtained, namely precision, recall accuracy, and F1-score as in Table 4 and Figure 2:

Table 3. Confusion matrix of NB algorithm before and after using the IG feature and AdaBoost features

Metrix	Before	After
TP	82	85
FP	20	16
FN	25	23
TN	73	76

Table 4. Evaluation of the performance of NB classifier

Evaluation	NB	NB+ IG + AdaBoost
Accuracy	77.5	80.5
Precision	80.3921569	84.158416
Recall	76.635514	78.703704
F1-score	78.4688995	81.339713

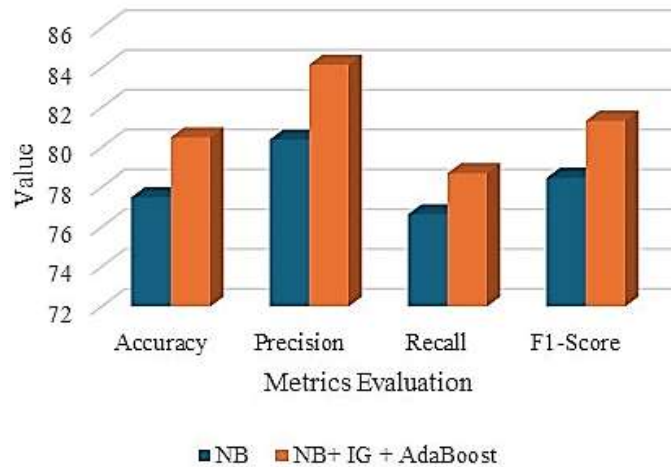


Figure 2. Evaluation of the performance of NB classifier

Table 5 presents the confusion matrix results of the KNN method before and after using the IG feature and the AdaBoost feature. Before using the feature, you can see that the composition results are true positive = 89, false positive = 21, false negative = 20, and true negative = 70. If you look at these results, the false positive and false negative values are still high. This happens because KNN is a very simple method and the KNN evaluation results are very dependent on feature scaling. Meanwhile, after using the IG feature and the AdaBoost feature, you can see the results of the composition true positive = 91, false positive = 14, false negative = 18, and true negative = 77. If you look at these results, the false positive, and false negative values have decreased compared to the previous FP and FN. The most significant decrease in value was also found in False Positive, resulting in an increase in the classification performance value. In this process, the influence of IG and Adaboost can also be proven, but the accuracy results are not yet significant. Even though they have increased, IG and AdaBoost cannot yet be categorized as suitable features for the KNN method.

Table 5. Confusion matrix of KNN algorithm before and after using the IG feature and AdaBoost features

Metrix	Before	After
TP	89	91
FP	21	14
FN	20	18
TN	70	77

After testing KNN with the IG and AdaBoost features, to see an illustration of the differences in the results, you can see in Table 6 and Figure 3. From Table 6 and Figure 3 you can see that the results of the KNN test without using the IG and AdaBoost features obtained an accuracy value of 79.5 %; precision 80.91%; recall of 81.65%; and F1-score of 81.27%. After using IG and AdaBoost, the accuracy value increased by 84%; precision to 86.67%; recall became 83.48%; and F1-score became 85.04%. This means that the KNN algorithm experienced an increase in accuracy of 4.5%, precision of 5.75%, recall of 1.83%, and F1-score of 3.76%. From the confusion matrix above, each classification performance evaluation value is obtained, namely precision, recall accuracy, and F1-score as in Table 6 and Figure 3.

Table 6. Evaluation of the performance of KNN classifier

Evaluation	KNN	KNN + IG + AdaBoost
Accuracy	79.5	84
Precision	80.90909	86.666667
Recall	81.65138	83.486239
F1-score	81.27854	85.046729

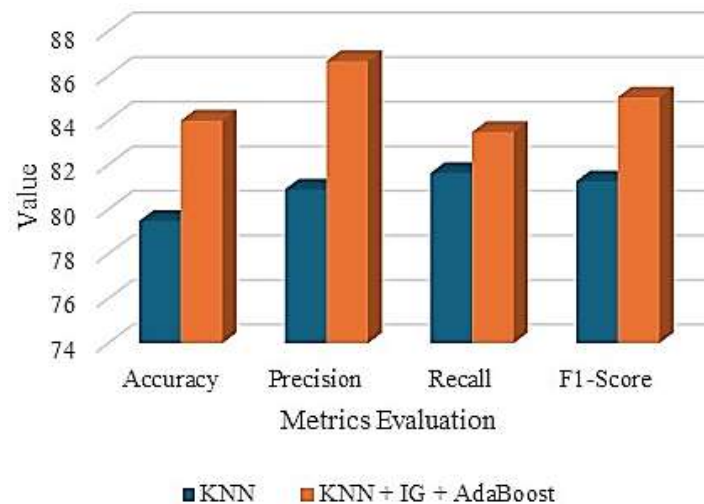


Figure 3. Evaluation of the performance of KNN classifier

Table 7 presents before and after using the IG feature and the AdaBoost feature. Before using the feature, you can see that the composition results are true positive = 79, false positive = 29, false negative = 22, and true negative = 70. If you look at these results, the False Positive and false negative values are still high. This happens because RF is an algorithm that requires more computing resources. In this case, RF also takes a long time to carry out classification. Meanwhile, after using the IG feature and the AdaBoost feature. You can see the composition results of true positive = 95, false positive = 16, false negative = 15, and true negative = 74. If you look at these results, the false positive, and false negative values have decreased compared to the previous FP and FN. The most significant decrease in value was also found in false positive, causing the classification performance value to increase. In this process, the influence of IG and AdaBoost can be proven, and the classification performance evaluation results also experience a significant increase. This significant increase in value is caused by the IG feature. This IG is suitable when combined with the RF algorithm.

Table 7. Confusion matrix of RF algorithm before and after using the IG feature and AdaBoost features

Metrix	Before	After
TP	79	95
FP	29	16
FN	22	15
TN	70	74

From the confusion matrix above, each classification performance evaluation value is obtained, namely precision, recall accuracy, and F1-score as in Table 8 and Figure 4. From Table 8 and Figure 4, it can be seen that the results of RF testing without using the IG and AdaBoost features obtained an accuracy value of 74.5%; precision of 73.14%; recall of 78.21%; and F1-score of 75.59%. After using IG and AdaBoost, the accuracy value increased by 84.5%; precision to 85.58%; recall was 86.36%; and F1-score was 85.97%. This means that the RF algorithm experienced an increase in accuracy of 10%, precision of 12.43%, recall of 8.14%, and F1-score of 10.37%.

Table 8. Evaluation of the performance of RF classifier

Evaluation	RF	RF + IG + AdaBoost
Accuracy	74.5	84.5
Precision	73.14815	85.58586
Recall	78.21782	86.363636
F1-score	75.59809	85.972851

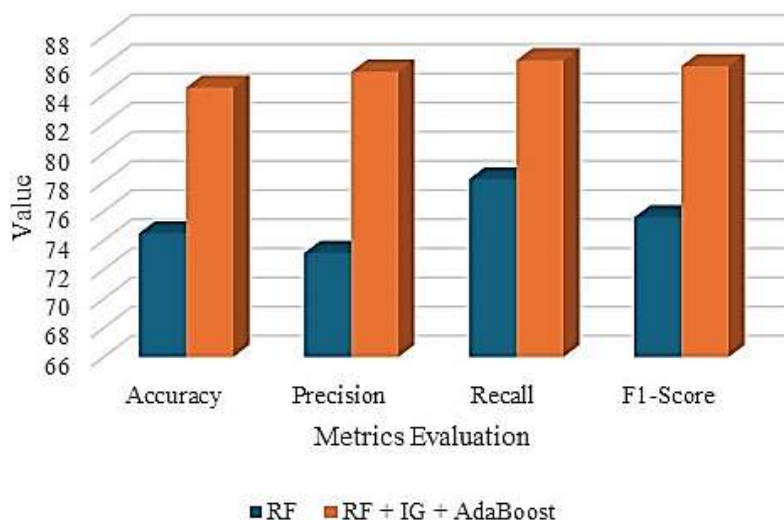


Figure 4. Evaluation of the performance of RF classifier

Based on this experiment, we have succeeded in proving the effect of IG and AdaBoost in increasing the evaluation value of classification performance for several machine learning methods such as NB, KNN, and RF. From the discussion above, it can be seen that the results of the comparison of the three methods above, the three methods above experienced an increase after using IG and AdaBoost. NB and KNN have increased, but the increase is not very high, and the overall value is not evenly distributed. The recall precision and F1-score accuracy values experience quite high differences, this occurs because the difference in false positive and false negative values is too significant. In contrast to the RF algorithm, this algorithm experienced very significant results, the results of the RF classification performance evaluation also tended to be stable and almost evenly distributed between the accuracy, precision recall, and F1-score values. This is because the false positive and false negative values are almost evenly distributed, another reason is because one of the supporting features, namely the IG feature, is very compatible with the RF algorithm.

4. CONCLUSION





In this paper, several machine learning methods such as NB, KNN, and RF as well as feature selection such as IG and AdaBoost are used to determine the effect of increasing the evaluation value of classification performance on restaurant review data. The experiment compares several previous and later machine learning methods using feature selection. Based on tests that have been carried out on several machine learning methods and a combination of the two selection features, all machine learning methods have experienced an increase in classification performance evaluation values, namely accuracy, precision, recall, and F1-score. But the best classifier is the RF algorithm. RF produces a significant increase in value after using the IG and AdaBoost features. RF also produces even accuracy, precision, recall, and F1-score values after using IG and AdaBoost. The results of evaluating the performance of the RF classification also tend to be stable and almost evenly distributed between the accuracy, precision recall, and F1-score values. one of the reasons is that the false positive and false negative values are almost evenly distributed, another cause is because one of the supporting features, namely the IG feature, is very compatible with the DT algorithm, of which RF is part of the DT.

REFERENCES





- [1] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, no. c, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.
- [2] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: capitalizing on big data," *Journal of Travel Research*, vol. 58, no. 2, pp. 175–191, 2019, doi: 10.1177/0047287517747753.
- [3] J. Jasmir, S. Nurmaini, R. F. Malik, and B. Tutuko, "Bigram feature extraction and conditional random fields model to improve text classification clinical trial document," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, pp. 886–892, 2021, doi: 10.12928/TELKOMNIKA.v19i3.18357.
- [4] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Applied Soft Computing*, vol. 86, 2019.
- [5] S. Nurmaini, R. U. Partan, W. Caesarendra, and T. Dewi, "An automated ECG beat classification system using deep neural networks with an unsupervised feature extraction technique," *Applied Sciences*, vol. 9, 2019.

- [6] M. S. Rahman and H. Reza, "A systematic review towards big data analytics in social media," *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 228–244, 2022, doi: 10.26599/BDMA.2022.9020009.
- [7] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," *International Journal of Scientific and Technology Research*, vol. 9, no. 2, pp. 601–609, 2020.
- [8] E. H. Muktafin and P. Kusriani, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 1, pp. 146–154, 2021, doi: 10.12928/TELKOMNIKA.V19I1.17417.
- [9] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.
- [10] E. T. Khalid, E. B. Talal, M. K. Faraj, and A. A. Yassin, "Sentiment analysis system for COVID-19 vaccinations using data of Twitter," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 2, pp. 1156–1164, 2022, doi: 10.11591/ijeecs.v26.i2.pp1156-1164.
- [11] Y. Wang and L. Feng, "Improved AdaBoost algorithm for classification based on noise confidence degree and weighted feature selection," *IEEE Access*, vol. 8, pp. 153011–153026, 2020, doi: 10.1109/ACCESS.2020.3017164.
- [12] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with IG for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.
- [13] K. Sherratt *et al.*, "Characterising IG s and losses when collecting multiple epidemic model outputs," *Epidemics*, p. 100765, 2024, doi: 10.1016/j.epidem.2024.100765.
- [14] N. M. G. D. Purnamasari, M. A. Fauzi, Indriati, and L. S. Dewi, "Cyberbullying identification in twitter using support vector machine and IG based feature selection," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 18, no. 3, pp. 1494–1500, 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.
- [15] W. Wang and D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," *Information Sciences*, vol. 563, pp. 358–374, 2021, doi: 10.1016/j.ins.2021.03.042.
- [16] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, "Ensemble machine learning model for classification of spam product reviews," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8857570.
- [17] H. Kim, J. Kim, J. Kim, and P. Lim, "Towards perfect text classification with Wikipedia-based semantic NB learning," *Neurocomputing*, pp. 1–7, 2018, doi: 10.1016/j.neucom.2018.07.002.
- [18] J. Jasmir, S. Nurmaini, and B. Tutuko, "Fine-grained algorithm for improving knn computational performance on clinical trials text classification," *Big Data and Cognitive Computing*, vol. 5, no. 4, 2021, doi: 10.3390/bdcc5040060.
- [19] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, "Feature extraction based text classification using K-nearest neighbor algorithm," *IJCSNS International Journal of Computer Science and Network Security*, vol. 18, no. 12, pp. 95–101, 2018.
- [20] T. Salles, M. Gonçalves, V. Rodrigues, and L. Rocha, "Improving random forests by neighborhood projection for effective text classification," *Information Systems*, vol. 77, pp. 1–21, 2018, doi: 10.1016/j.is.2018.05.006.
- [21] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2733–2742, 2022, doi: 10.1016/j.jksuci.2022.03.012.
- [22] "Restaurant customer reviews," <https://www.kaggle.com/datasets/vigneshwarsofficial/reviews/versions/1?resource=download>.
- [23] N. Ning and Y. Tang, "Evaluation of an information flow gain algorithm for microsensor information flow in limber motor rehabilitation," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6638038.
- [24] S. Xu, Y. Liang, Y. Li, S. S. Du, and Y. Wu, "Beyond IG: an empirical benchmark for low-switching-cost reinforcement learning," pp. 1–19, 2023.
- [25] O. Chamorro-Atalaya *et al.*, "Supervised learning through k-nearest neighbor, used in the prediction of university teaching performance," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 3, pp. 1625–1634, 2022, doi: 10.11591/ijeecs.v27.i3.pp1625-1634.
- [26] B. Y. Pratama, "Personality classification based on twitter text using Naive Bayes, KNN and SVM," pp. 170–174, 2015.
- [27] O. Anava and K. Y. Levy, "K-nearest neighbors: from global to local," *Advances in Neural Information Processing Systems*, no. Nips, pp. 4923–4931, 2016.
- [28] Y. Yang and X. Liu, "A re-examination of text categorization methods," *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49, 1999.
- [29] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, no. 2018, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.
- [30] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in Naive Bayes for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2508–2521, 2016, doi: 10.1109/TKDE.2016.2563436.
- [31] M. Elhadad, D. Gabay, and Y. Netzer, "Automatic evaluation of search ontologies in the entertainment domain using text classification," *Applied Semantic Technologies: Using Semantics in Intelligent Information Processing*, pp. 351–367, 2011.
- [32] A. Choi, N. Tavabi, and A. Darwiche, "Structured structured features in Naive Bayes classification," *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 3233–3240, 2016.
- [33] K. U. Santoshi, S. S. Bhavya, Y. B. Sri, and B. Venkateswarlu, "Twitter spam detection using NB classifier," *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, pp. 773–777, 2021, doi: 10.1109/ICICT50816.2021.9358579.
- [34] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.
- [35] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2016.





BIOGRAPHIES OF AUTHORS

Jasmir Jasmir     is senior lecture at Universitas Dinamika Bangsa Jambi, Indonesia. He received his Bachelor in Computer Engineering in 1995 and Master degree in Information Technology in 2006 from Universitas Putra Indonesia YPTK Padang, Indonesia. He receives a Doctor in Informatics Engineering at Universitas Sriwijaya Palembang, Indonesia in 2022. His research interest is data mining, machine learning and deep learning for natural language processing and its application. He can be contacted at email: ijay_jasmir@yahoo.com.







Dodo Zaenal Abidin     is senior lecture at Universitas Dinamika Bangsa Jambi, Indonesia. He received his Bachelor in Information System in Universitas Nurdin Hamzah Jambi in 2000 and Master degree in Information Technology in Universitas Putra Indonesia YPTK Padang, Indonesia in 2008. He receives a Doctor in Informatics Engineering at Universitas Sriwijaya Palembang, Indonesia in 2022. His research interest is data base, signal processing, and natural language processing. He can be contacted at email: dodozaenalabidin@gmail.com.



Fachruddin Fachruddin     received a Bachelor's degree (S.Pt) in Agriculture from Jambi University in 1998. He obtained a Master's degree (M.S.I) in Information Systems from the Dinamika Bangsa University in 2011. He receives a Doctor in Informatics Engineering at Universitas Sriwijaya Palembang, Indonesia in 2022. He is a Lecturer in Computer Science, Informatics Engineering, Universitas Dinamika Bangsa (UNAMA). Artificial intelligence and information systems. He can be contacted at email: fachruddin.stikom@gmail.com.



Willy Riyadi     completed his Bachelor of Computer Systems in 2011 and completed his Masters in Information Systems in 2014 in Universitas Dinamika Bangsa. He is a lecturer at Universitas Dinamika Bangsa, Jambi, specializing in system control, artificial intelligence, and machine learning. He can be contacted at email: wriyadi5@gmail.com.