Scientific Journal of Informatics
Vol. 8, No. 2, November 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning

Nurul Hidayat[1]*, M. Faris Al Hakim[2], Jumanto[3]
[1]Informatics Department, Faculty of Engineering,
Universitas Jenderal Soedirman, Indonesia
[2,3] Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: Halal tourism or Muslim-friendly tourism has significant potential for the tourism industry in Indonesia. According to Cresent Rating, the world's leading authority on halal-friendly travel, one of the indicators for halal tourism is the availability of choices for halal foods. To support halal tourism, unfortunately, not all restaurants around the tourism object or in the city where the tourism object is located have labels or information that easily makes people know about halal food in the restaurant.
Methods/Study design/approach: The data in this research was obtained from online media such as Google Maps, TripAdvisor, and Zoomato. The data consists of 870 data with the classification of halal food restaurants and 590 data with the reverse classification. Machine learning methods were chosen as classifiers. Some of them were Naive Bayes, Support Vector Machine, and K-Nearest Neighbor.
Result/Findings: This research shows that the proposed method achieved an accuracy of 95,9% for Support Vector Machine, 93,8% for Multinomial Naive Bayes, and 91% for K-Nearest Neighbor. In the future, our result will be to support the halal tourism environment in terms of technology.
Novelty/Originality/Value: In this study, we utilize restaurant reviews done by visitors to get information about the classification of halal food restaurants.

Keywords: Halal Food, Classification, Support Vector Machine, Multinomial Naive Bayes, K-Nearest Neighbor

Received October 2021 / **Revised** November 2021 / **Accepted** November 2021

INTRODUCTION

Halal tourism or Muslim-friendly tourism generally is part of the tourism industry for Muslim tourists. From 2000 until 2020, Global Muslim Travel Index (GMTI) calculates that the number of Muslim world tourists continues to grow 27 percent per year and be predicted that reach 158 million. This growth rate far exceeds the growth of world tourists, which is only 6.4 percent per year, according to WTTC records. In 2019, Indonesia was awarded by GMTI as one of the best world halal tourism destinations. The conditions show that halal tourism has significant potential for the tourism industry in Indonesia. According to Cresent Rating, the world's leading authority on halal-friendly travel, one of the indicators for halal tourism is the availability of choices for halal foods. Unfortunately, to support halal tourism, not all restaurants around the tourism object or in the city where the tourism object is located have labels or information that makes people quickly know about halal food in the restaurant.

In this era, many customers share their opinions as reviews about their experience in online media, such as Google Maps, TripAdvisor, and Zoomato. Those reviews can be utilized for classification in text. Text classification is a model construction problem that can classify new documents into pre-defined classes [1][1], [2]. As a classification result, those reviews will be categorized as halal or non-halal. In that case, text classification will uncover which restaurant be consistent for halal food and not.

[1]*Corresponding author.
Email addresses: nurulhidayat@mail.unsoed.ac.id (hidayat), farishakim@mail.unnes.ac.id (hakim), jumanto@mail.unnes.ac.id (jumanto)
DOI: 10.15294/sji.v8i2.25356

Some algorithms can be used for text classification. Some of them are K-Nearest Neighbor, Naive Bayes, Decision Trees, Support Vector Machine, and The Ensemble Learning technique [3]–[5]. [7] used Naive Bayes Classifier to do sentiment analysis for food culinary at Tripadvisor website with an accuracy of 98,67%. In 2019, Londo et al. (2019) [8] proposed text classification focused on Indonesian News Articles as a study. As a result, Support Vector Machine generates 93% for each metric performance as the best model. Suharno et al. (2017) [9] used K-Nearest Neighbor to classify online complaint documents in Indonesian. The result showed that K-Nearest Neighbor could generate 78% as the highest value of *f-measure*.

Multinomial Naive Bayes, Support Vector Machine, and K-Nearest Neighbor were the best three machine learning algorithms for text classification. Thus, this research intends to find machine learning methods that generate the best result to accurately classify halal food restaurants based on restaurant reviews in Indonesian language.

METHODS

This section describes several steps used in this research. The research was begun with data acquisition and was followed by data pre-processing, feature selection, classification, and results. Figure 1 below shows several steps in this research.

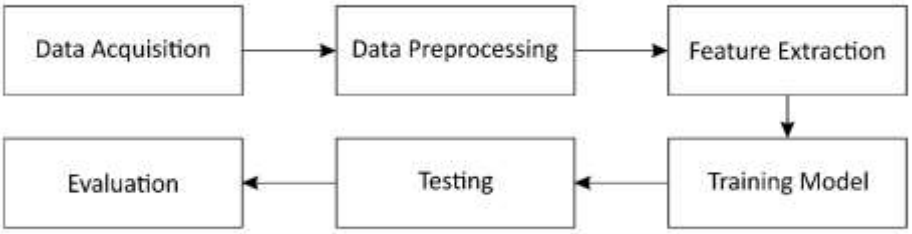

Figure 1. Research Methodology

Data Acquisition

In the data acquisition step, collecting and labeling data were done. Data used in this research was restaurant reviews from several sites such as Google Maps, TripAdvisor, and Zomato. Data retrieval was done using the WebHarvy tool. In this research, the reviews publicized on several sites are crawled with a total number of 1460 reviews. Every review was labeled as a halal food restaurant and non-halal food restaurant. Label 1 was given to a restaurant that only provided halal food, and label 0 was given to a restaurant that provided non-halal (haram) food. Although the restaurant provided halal and haram food, it was labeled with 0. These reviews consist of 870 reviews for the halal category and 590 reviews for the non-halal category.

Data Pre-processing

This phase aimed to clean and prepare data that data was ready to be used. Some processes done in this phase were case conversion, cleaning items, tokenizing, stopword removal, and stemming. Case conversion made all letters uniform by changing letters to lowercase like "BagUs" became "bagus". Cleaning items were done by removing characters and punctuation that often appeared and did not have much meaning, like "-, @, #, ?, !". Tokenizing was a process to split text based on composer words. The process of removing words that have less meaning in languages, such as prepositions and conjunctions (example: dan, dimana, di, yang, etc.) was called stopword removal. The last in this phase was stemming, which aimed to get the base form of each word like "berlari" and "melarikan" became "lari".

Feature Extraction

Feature Extraction in this research was done using TF-IDF (Term Frequency-Inverse Document Frequency) that was TF and IDF as the combination. TF-IDF method had a direct impact on the accuracy and speed of the learning algorithm [10]. The equation below shows TF-IDF that contains The representation of the weight of a term in a document:

$$W(d,t) = TF(d,t) * log\ (\frac{N}{df(t)}) \qquad (1)$$

Here $TF(d,t)$ is the value of the word d in term t. The number of documents represents by N and df(t) represents the number of documents containing the term t. The main idea of weighing in IDF is that t is a feature to distinguish the documents from one another, where not many documents contain t.

Classification

After the data pre-processing and feature extraction step, classification was the next step to do. In this step, training and testing data were done. Extracted features from the reviews were used to train models that classified reviews into 1 as halal food restaurant and 0 as not halal food restaurant. This research used Naive Bayes, KNN, and Support Vector Machine as classifiers.

Naive Bayes Classifier. This classifier has been one of the most used methods for text classification since the 1950s. Naive Bayes, also known as the Bayes theorem, uses probability and opportunity approaches. In this research, Multinomial Naive Bayes was chosen as the representation of the Naive Bayes Classifier. Multinomial naive bayes [11] work competently when multiple emergences of the words have a crucial impact on the classification problem. It finds the probability of the review that belongs to a particular class by combining the probability of all features in the class. It can be defined by the bayes theorem, where all the variables in a given class C are conditionally independent of each other. Naive Bayes algorithm can be described as follows:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \quad (2)$$

Where d is a document and c indicates classes.

$$\hat{y} = arg\,max\, P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (3)$$

K-Nearest Neighbor. The K-Nearest Neighbors algorithm (KNN) is a non-parametric classification technique. One of the method's usability has been implemented for text classification in many research areas [12] in the past decades. Like the name, the basic concept of KNN is to find its nearest neighbors as many ask among all the data in the training set and score category candidates based on the class of k-neighbors. The score was given to every neighbor from the similarity with each neighbor. The highest score will be assigned as its class. KNN can be formulated as follow:

$$f(x) = arg\,\max_{j} S(x, C_j) \quad (4)$$

$$= \sum_{d_i \in KNN} sim(x, d_i) y(d_i, C_j) \quad (5)$$

where S refers to the score value concerning for $S(x, Cj)$, the score value of candidate i to a class of j, and the output of $f(x)$ is a label to the test set document.

Support Vector Machine. The Support Vector Machine algorithm (SVM) was invented by Vapnik and Chervonenkis [13] in 1963 as the original maximum-margin hyperplane algorithm constructed a linear classifier. In 1992, Boser et al. [14] conformed this type into a nonlinear formulation. For text classification context, let $x1, x2, \ldots, xl$ be training examples belonging to one class *X*, where *X* is a compact subset of R^N [15]. Binary classifier used in this research can be formulated as follow:

$$min\,\frac{1}{2}||w||^2 + \frac{1}{vl}\sum_{i=1}^{l} \xi_i - \rho \quad (6)$$

subject to:

$$\big(w.\Phi(x)\big) \geq \rho - \xi_i \quad i = 1, 2, \ldots, l \quad \xi_i \geq 0 \quad (7)$$

If this problem was solved by w and ρ, then the decision function is given by:

$$f(x) = sign(\big(w.\Phi(x)\big) - \rho \quad (8)$$

Evaluation Measures

An evaluation was done to know the performance of the model built. The classification model was assessed using standard measures such as accuracy, precision, and recall. The confusion matrix was chosen to represent the classification result shown in Table 1.

Table 1. Confusion matrix

Classification		Classified as	
		Positive	Negative
	+	True positive	False-positive
	-	False-negative	True negative

The following is a brief explanation of Table 1.

- True Positive (TP): positive review document is classified correctly
- True Negative (TN): negative review document is classified correctly
- False Positive (FP): positive review document is classified as negative
- False Negative (FN): negative review document is classified as positive

Precision refers to a positive predictive value. It has a high ability to retrieve reviews that are judged by the user as being relevant. The value of precision can be generated from the following equation:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall shows the average success of the model in the process of finding information. It can be obtained by dividing between correct classified positive reviews and all positive reviews as in the following equation:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

The combination of Recall and Precision called F_1 by Van Rijsbergen (1979) was used to evaluate the classification effectiveness. It can be defined mathematically as:

$$F_1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (11)$$

Accuracy is the degree of truth between the predictive value and the actual value. Accuracy showed the overall accuracy of the result of the model classification. It is generated by dividing between the sum of all numbers predicted correctly, and all of the data like the equation follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

RESULT AND DISCUSSION

This section shows the explanation of some experiments results. Based on 1460 reviews that consist of 870 data classified as halal food restaurants and 590 data classified as non-halal food restaurants and with cross-validation where the comparison between training and testing data used is 80:20, the results are explained as follows.

The first method, Multinomial Naive Bayes, generated accuracy, which is around 93,8%, Precision in 91%, 99,4% for Recall, and 95% in F1-Score. Table 2 shows the result of Multinomial Naive Bayes for each class. There is a small number of data on a non-halal class that is misclassified by this method. It is also supported by[16][17] that naive Bayes can perform classification well.

Table 2. Result of Multinomial Naive Bayes

Classification	Precision	Recall	F1-Score
Halal	0,91	0,99	0,95
Non-Halal	0,99	0,86	0,92

The second method, K-Nearest Neighbor generated accuracy, which is around 91%, Precision 91%, 94,2% in Recall, and 92,6% in F1-Score. Table 3 presents K-Nearest Neighbor for each class. This method has the smallest score if compared with the others. The performance metrics score achieved just 92-93%. Based on

the confusion matrix, 6 data are misclassified by it include 10 data in halal class and 16 data in non-halal class.

Table 3. Result of K-Nearest Neighbor

Classification	Precision	Recall	F1-Score
Halal	0,91	0,94	0,93
Non-Halal	0,91	0,87	0,89

The last method, Support Vector Machine, yielded accuracy, which is around 95,9%, Precision 95%, 98,3% for Recall, and 96,6% for F1-Score. Table 4 reported the result of the Support Vector Machine for each class. This method also still misclassified a small number of data in the non-halal category.

Table 4. Result of Support Vector Machine

Classification	Precision	Recall	F1-Score
Halal	0,95	0,92	0,95
Non-Halal	0,97	0,98	0,97

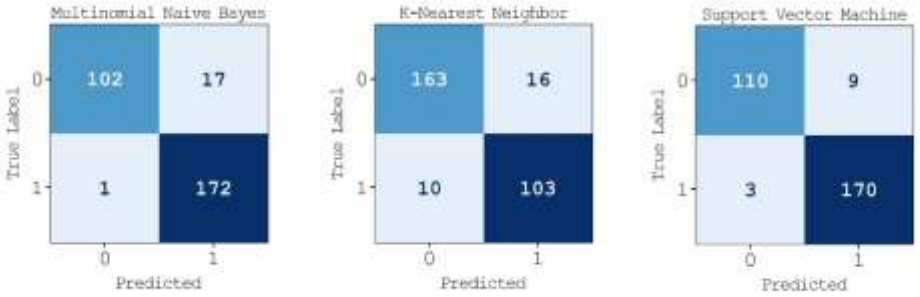

Figure 2. Confusion Matrix Evaluation

It can be seen in figure 2, confusion matrix evaluation, Support Vector Machine classified 280 data correctly, and 12 data were not accurate. Its result is better than Multinomial Naive Bayes that was classified 274 data true, and K-Nearest Neighbor that was classified 266. In the non-halal class, Support Vector Machine was also classified better than the two other classes, but it was different in the halal class. Multinomial Naive Bayes only generated 1 data false in halal class; Support Vector Machine generated 2 more data false than it. K-Nearest Neighbor produced the highest value of false classification with 10 data in halal class and 16 data in non-halal class.

All classifiers yielded higher misclassified in the non-halal category than in the halal category. Multinomial Naive Bayes failed to predict 17 data in non-halal class, or 16 data more than in halal class. K-Nearest Neighbor generated 6 misclassified data more in a non-halal class than in a halal class with a total of misclassified data is 26 data. Support Vector Machine that generated the smallest misclassified data, only 12 data was classified false, also misclassified higher in non-halal class with value 9 and 3 for halal class. One of the factors that led to the condition was not too many words representing haram food in the data obtained. The condition was also strengthened by not many restaurants that provide non-halal food on the Tripadvisor site.

Table 5. Classification Performance Result

Method	Precision	Recall	F1-Score	Accuracy
Multinomial Naive Bayes	91%	99,4%	95%	93,8%
K-Nearest Neighbor	91%	94,2%	92,6%	91%
Support Vector Machine	95%	98,3%	96,6%	95,9%

As to be shown in Table 5, Support Vector Machine provides the highest value of precision, f1-score, and accuracy with values 95%, 96,6%, and 95,9%, respectively. At the same time, the highest value of recall is

yielded by Multinomial Naive Bayes with a value of 99,4%. The total number of data, both training and testing data, becomes one of the factors that can affect the result—the higher number of data, the more various the data content.

CONCLUSION

Based on the experiments using machine learning, the result shows that the algorithm proposed can classify restaurant reviews in Indonesian as halal food restaurants and non-halal food restaurants. The results also indicate that Support Vector Machine had better performance than the 2 other methods. Support Vector Machine achieved a value of 95,9% of accuracy, Multinomial Naive Bayes as the second whose accuracy of 93,8%, and K-Nearest Neighbor as the last whose accuracy 91%. In the future, other types of weighting methods can be used to improve performance. Increasing various content and the number of data will be a challenge to be done to get better support in data. Other powerful methods also can be an alternative comparison to generate results optimally.

REFERENCES

[1] B. Liu, *Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data*. New Jersey: Springer Heidelberg, 2006.

[2] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (Data-Centric Systems and Applications)*. New Jersey: Springer Verlag, 2006.

[3] C. D. Manning, R. Prabhakar, and S. Hinrich, *Introduction to Information Retrieval*. Cambridge, UK: Cambride University Press, 2008.

[4] K. Aas and L. Eikvil, "Text Categorisation: A survey. Technical Report," vol. 45, 1999.

[5] T. S. Guzella and W. M. Caminhas, "A Review of Machine Learning Approaches to Spam Filtering," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10206–20222, 2009.

[6] C. C. Aggarwal and C. Zhai, *A Survey of Text Classification Algorithms. In Mining Text Data*. Berlin: Springer, 2012.

[7] S. A. Azzahra and A. Wibowo, "Analisis Sentimen Multi-Aspek Berbasis Konversi Ikon Emosi dengan Algoritme Naïve Bayes untuk Ulasan Wisata Kuliner Pada Web Tripadvisor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 4, p. 737, 2020.

[8] G. L. Yovellia Londo, D. H. Kartawijaya, H. T. Ivariyani, P. W. P. Yohanes Sigit, A. P. Muhammad Rafi, and D. Ariyandi, "A Study of Text Classification for Indonesian News Article," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 205–208, 2019.

[9] C. F. Suharno, M. A. Fauzi, and R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-," *Syst. Inf. Syst. Informatics J.*, vol. 03, no. 01, pp. 25–32, 2017.

[10] N. Rezaeian and G. Novikova, "Persian text classification using naive bayes algorithms and support vector machine algorithm," *Indones. J. Electr. Eng. Informatics*, vol. 8, no. 1, pp. 178–188, 2020.

[11] J.-Y. Yoo and D. Yang, "Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier," *Adv. Sci. Technol. Lett.*, vol. 111, no. Computer and Computing Science, pp. 263–266, 2015.

[12] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1503–1509, 2012.

[13] V. Vapnik and A. Y. Chervonenkis, "A Class of Algorithms for Pattern Recognition Learning," *Avtomat. i Telemekh*, vol. 25, no. 6, pp. 937–945, 1964.

[14] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *The Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[15] L. M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, 2001.

[16] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, 2021.

[17] O. Arifin, K. Saputra, and H. Fathoni, "Implementation of Data Mining using Naïve Bayes Classifier Method in Food Crop Prediction," *Sci. J. Informatics*, vol. 8, no. 1, pp. 43–50, 2021.