INTERNATIONAL JOURNAL ON FOOD

Halal Food Prediction Using the Similarity Graph Algorithms

Nur Aini Rakhmawati[ab], Girraz Karyo Utomo[a] , Rarasmaya Indraswari[a], and Irfan Rifqi Susetyo[a]

[a]*Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.*
[b]*Halal Center, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
nur.aini@is.its.ac.id, utomogirraz@gmail.com, raras@its.ac.id, irfnrifqi@gmail.com

ABSTRACT

Only halal food is food allowed by Islamic sharia. In contrast, haram food is forbidden, for example, alcohol, pork, blood, carrion, and meat not slaughtered according to sharia regulations. In Indonesia, halal certificates are issued by the Halal Product Guarantee Agency (BPJPH) based on Article 39 of Law Number 33/2014 concerning Halal Product Guarantees. A halal certificate guarantees that food is composed of halal ingredients. However, many halal food products do not have a halal certificate. Therefore, it is useful to estimate the halal status of food products that have not been certified. In this work, we attempted to predict the halal status of food using several graph similarity algorithms. We acquired product data from the KlikIndomaret website, which contained the food product name, the composition of the food product, and the manufacturer. Then, we crawled the halal food database on the Halal MUI website. Both datasets were merged into a single dataset based on the product names. Then, similarity algorithms such as Jaccard Similarity, Approximate Nearest Neighbor, Adamic Adar and Preferential Attachment were performed on the products in the dataset. The accuracy of each algorithm evaluated by F-measure.

Keywords: halal food; similarity algorithms; food products; graph algorithm

1 Introduction

Halal food is any food that meets Islamic sharia regulations (Rezai et al., 2012), while haram food is any food that is not allowed to be consumed, such as alcohol, pork, blood, carrion, or meat not slaughtered according to sharia regulations (Salman and Kamran Siddiqui, 2011). Therefore, Muslims must always be concerned about what is in their food. Like most other Muslim countries, Indonesia issues halal certificates for food products, through the Halal Product Assurance Organizing Agency (BPJPH) (Mukidi, 2020; Svinarky and Malau, 2020). The number of halal-certified products increases year by year (Warto and Samsuri, 2020).

The Global Islamic Economy (GIE) report for 2020/2021 shows that Indonesia rose to rank 4 after being ranked fifth and tenth in the respective previous periods (2019/2020) and (2018/2019) (State of the Global Islamic Economy, 2021). The improvement of Indonesia's ranking is a positive result of the issuance of the 2019-2024 Indonesian Islamic Economic Master Plan issued by the Ministry of National Development Planning/National Development Planning Agency in collaboration with the National Sharia Finance Committee (KNKS) (Ministry of National Development Planning and Committee, 2018). However, Indonesia is not in the top-ten in the categories of halal food, media and recreation, medicine, and cosmetics. Only ten percent of products are halal-certified in Indonesia (Petriella, 2019). Meanwhile, Indonesia is the country with the largest Muslim population globally.

Halal food is a primary need for every Muslim in the world. GIE predicts that on a global scale, Muslims spend two trillion US dollars on halal food annually (State of the Global Islamic Economy, 2021). As a result, there is a high demand for halal certification. To deal with the lack of halal certification, the halal status of products could be detected by looking at the ingredients (Karimah and Darwanto, 2021). One of the most famous rules is the absence of any pork ingredients (Çetin and Dinçer, 2016). There are several other products and their derivatives that are not allowed to be consumed, such as blood, dog meat, reptile meat, alcohol, etc. (Jia and Chaozhi, 2021). Moreover, some food additives are from haram sources (Gültekin et al., 2020). It is hard to detect the halal status of food if the ingredients are not well known, as in the case of food additives. Rakhmawati et al. (2018) predicted the halal status of non-halal-certified products using Euclidian distance and cosine similarity. Their method calculates the similarity of ingredients in these products with those in products listed in the Linked Open Data (LOD) Halal dataset (Rakhmawati et al., 2021). If a non-halal-certified product has a similarity value greater than 0.8 with a halal-certified product, the product can be seen as halal and recommended for consumption by Muslims.

Food prediction and recommendation research have been conducted with various methods. Petković et al. (2021) created DIETHUB, a tool for predicting food recipes based on diet activities. They used FoodBase (Popovski et al., 2019) as the dataset. The FoodBase corpus consists of 12,844 food entities, with 2,105 unique foods. The tool classifies the dataset into five categories: Appetizers and Snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. Ispirova et al. (2020) clustered the nutrient information from food consumption data in Slovenia. The dataset contained 3,265 food data. They combined word embedding, graph embedding, and machine learning to cluster the dataset into 23 categories. Yunus et al. (2019) employed different deep learning models to identify food from food images. Rakhmawati and Jannah (2021) calculated the similarity of food ingredients using several methods, such as Jaccard Similarity, Jaro-Winkler Distance, Levenshtein Distance, Wordnet LCH Similarity, and Fuzzy String Matching. The algorithms were performed on the Open Food Fact dataset (https://world.openfoodfacts.org/). Kautsar and Rakhmawati (2019) integrated data from seven halal certification bodies; 11.3% of the retrieved products already existed in the LOD Halal dataset. The best result was obtained by Levenshtein Similarity, with 2,685 products out of a total of 10,153 products. Rakhmawati and Najib (2020) investigated the similarity between food products and then created product interlinks in the LOD Halal dataset. Node2Vec (Grover and Leskovec, 2016) was implemented to measure the similarity between the products.

Currently, there are about 500 halal certification institutions worldwide. However, no international board manages a central database to integrate the halal certification data from these various institutions (Tieman and Williams, 2019). As a result, a single product may have multiple halal certificates from different institutions, since each institution has its own standards and regulations. Therefore, Rakhmawati et al. (2021) proposed LOD for halal food products. LOD uses data published on the web, where the data from multiple datasets are connected to each other. LOD is a knowledge graph construction so that a graph algorithm can be implemented on LOD. A knowledge graph is a great tool for integrating data from various sources (Rezaul Karim et al., 2019). We used several different graph algorithms in our work, i.e., Jaccard Similarity, Approximate Nearest Neighbor (ANN)(Dong et al., 2011), Adamic Adar (Adamic and Adar, 2003), and Preferential Attachment (Albert and Barabási, 2002). These algorithms first identify the similarity between halal products and then present the results in graph visualizations. These presentations can assist Muslims in identifying non-halal-certified products that are connected to halal-certified products. A non-halal-certified product is a product that has not been certified by any halal certification body.

The rest of this paper is divided into four sections: Section 2 describes the proposed methodology. Section 3 presents our experiment and the result. The paper is concluded in Section 4.

2   Methodology

Our methodology consists of four steps, namely data acquisition, similarity analysis, evaluation, and visualization (Fig 1). The four steps can be explained as follows.

2.1   Data Acquisition

KlikIndomaret (https://www.klikindomaret.com/) was the primary resource for our research. This website is an online shopping website from the Indomaret supermarket. The data retrieved concerned food and beverages and included the name of the product, the ingredients, and the manufacturer. Data retrieval was done using Scrapy Python (https://scrapy.org/). Scrapy is a Python library for extracting data from websites. There were 3,274 food and beverage products in 21 categories on the website.

We also collected data from the official Halal MUI website (Majelis Ulama Indonesia, https://www.halalmui.org/), which contains 401,650 products in 36 categories. These data served as complementary data for the Indomaret data since the halal status of the products is unavailable at KllikIndomaret.

The KlikIndomaret and MUI Halal datasets were merged into a single dataset using the Jaccard Similarity algorithm. We compared the product names and calculated the Jaccard similarity. If the similarity value was greater than 0.7, then the products in both datasets were considered similar.

The result of data integration was converted to a graph consisting of four types of nodes: ingredient, manufacturer, certificate, and food product. The graph dataset can be found at https://github.com/utomogirraz/Paper-English-Halal.
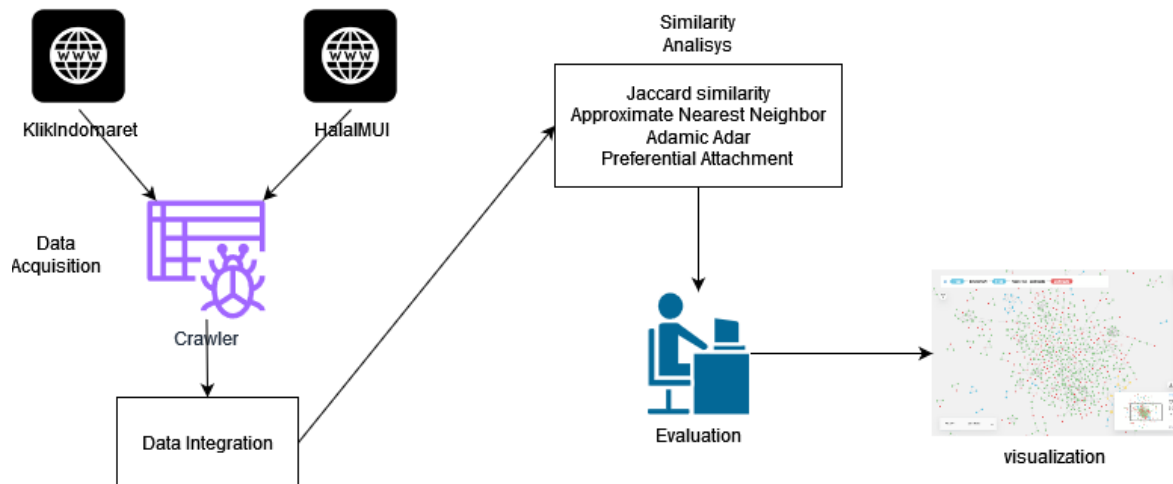

Figure 1. Methodology

2.2   Similarity Analysis

We performed the following graph similarity algorithms on the graph dataset:

- Jaccard Similarity (Rakhmawati and Jannah, 2021) calculates dataset similarity, dissimilarity, and distance. It measures proximity and checks data redundancy. Jaccard similarity is defined as the ratio of the number of common neighbors and the number of neighbors.

- Approximate Nearest Neighbor (ANN) (Dong et al., 2011) is an algorithm that builds a k-Nearest Neighbor graph model. This algorithm calculates the similarity based on Jaccard similarity, cosine similarity, Euclidean distance (Liu et al., 2019), or Pearson similarity (Benesty et al., 2009).

- Adamic Adar (Adamic and Adar, 2003) is a measurement used for computing the nearest node based on shared neighbors, where a node connected to all nodes gets the lowest value. The Adamic Adar algorithm was first introduced in 2003 by Adamic and Adar to predict linkages across social networks.

- Preferential Attachment (Albert and Barabási, 2002) calculates the nearest node by looking at its shared neighbors. In this algorithm, the more linked a node is, the more likely it is to get new connections.

Neo4J provides all four algorithms. Neo4J is a graph database platform that supports many features for data analysis. To check the accuracy of the algorithms, we conducted semi-automated data labeling in two steps. In the first step, Jaccard Similarity was run. Then, the accuracy of the pairs of products was checked manually in the

second step. The number of product pairs was 282,492. The average number of correctly predicted labels was 66%, where the threshold was 0.42. The results of the Jaccard labeling were then used as reference for the other algorithms. After getting the similarity values, we created interlinks between both products of a pair with a similarity value greater than 0.42.

2.3   Evaluation

We conducted a human evaluation for assessing the Jaccard Similarity and ANN algorithms. Three evaluators checked the list of product pairs. To reduce the number of assessments of product pairs, we selected 271 pairs of products with a similarity value higher than 0.8 for both the Jaccard Similarity and the ANN algorithm. We calculated the recall, precision and F-measure values. The recall is the ratio between the number of correctly predicted product pairs and the number of product pairs, while the precision is the ratio between the number of correctly predicted products pairs and the number of predicted products. The F-measure is the weighted average of precision and recall. We also calculated the Kappa coefficient (Kraemer, 2015) to evaluate the agreement between evaluators. This measurement calculates the degree of agreement between evaluators in predicting the similarity between two products.

2.4   Visualization

Neo4JBloom, one of the Neo4J tools for creating code-less visualizations, was used to visualize the relationships between the products from the previous stage. This visualization depicts food products that have similar halal relationships. Only 21,239 out of 71,423 products are presented in this visualization since these products contained at least one ingredient. To give better insight, we also calculated the page ranks (Hashemi et al., 2020) of the products. PageRank assigns a score based on the number of relationships. Three colors were used for the products. Green represents halal-certified products. Red represents halal certificate nodes. Nodes with a yellow color do not have a halal certificate but do have a relationship with a halal-certified product (green node). Therefore, the yellow nodes are predicted as halal products. The remaining nodes are blue. Figure 2 shows our visualization graph for ANN similarity. The nodes are distributed based on their similarity to other nodes. Some green nodes are located in the same area, which means they have same certificate number. Several blue nodes and green nodes are separated from the largest node community, which means they do not have similarity with any node in the largest node community.
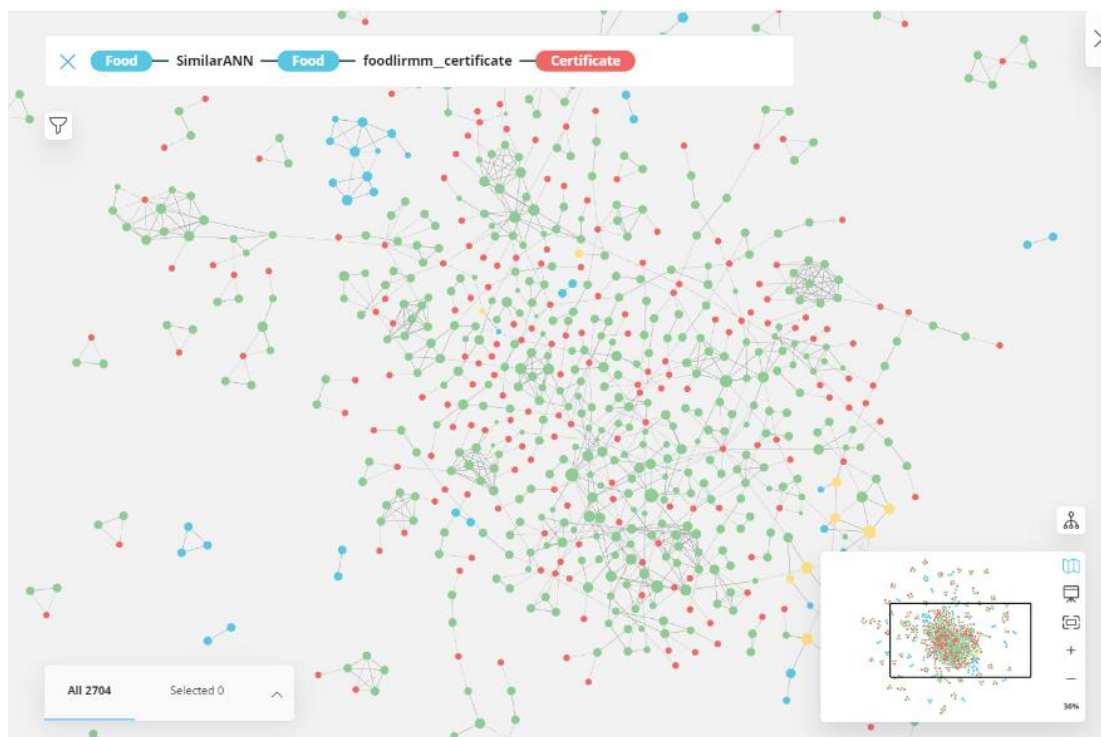

Figure 2. ANN Visualization Results

3 Results and Discussions

Table 1 describes the outcome of the human evaluation for checking the similarity of the products. Evaluators 1 and 2 classified the same numbers of products as similar and non-similar, respectively. In total, we had 182 agreements among the evaluators. Therefore, the coefficient kappa was 0.90, which means that the level of agreement between the evaluators was very strong.

Table 1.
Evaluator Assessment Results

Result	Similar	Not Similar
Evaluator 1	140	131
Evaluator 2	140	131
Evaluator 3	160	111

As can be seen in Table 2, the algorithm's accuracy was 99%, while the F-measure was 96%. It can be concluded that the ANN similarity algorithm could successfully predict the similarity of products.

Table 2.
Algorithm Measurement Value

Results	Value
Accuracy	0.9998
F-Measure	0.9645
Precision	1.0
Recall	0.9314

The number of nodes, relationships and certificate nodes generated by Jaccard Similarity, ANN and Preferential Attachment were precisely the same (Table 3). However, the average scores were different since the similarity calculations were different. The number of products nodes refers to the total number of product nodes that have a relationship with other product nodes, while the number of certificate nodes is the total number of halal certificate nodes owned by product nodes that have a relationship with other product nodes. The number of relationships is the sum of the relationships between product nodes that have similarity, while the number of certificate relationships is the sum of relationships between product nodes with a halal certificate and product nodes without a halal certificate. Adamic Adar produced fewer results than the other algorithms because it gives a lower score to more connected nodes.

Table 3.
Algorithm Comparison Results

Algorithm	Number of Product Nodes	Number of Certificate Nodes	Number of Relationships	Number of Certificate Relationships	Average Score of similarity algorithm
Jaccard	603	248	1332	521	0.621
ANN	603	248	1332	521	0.621
Adamic Adar	590	238	1279	508	2.803
Preferential Attachment	603	248	1332	521	152.67

Figure 3 shows a visualization of the products that do not have a relationship with a halal-certified product according to Jaccard Similarity. The coated peanuts (garlic powder) product has no relationship with *Garuda Pilus Pedas.*
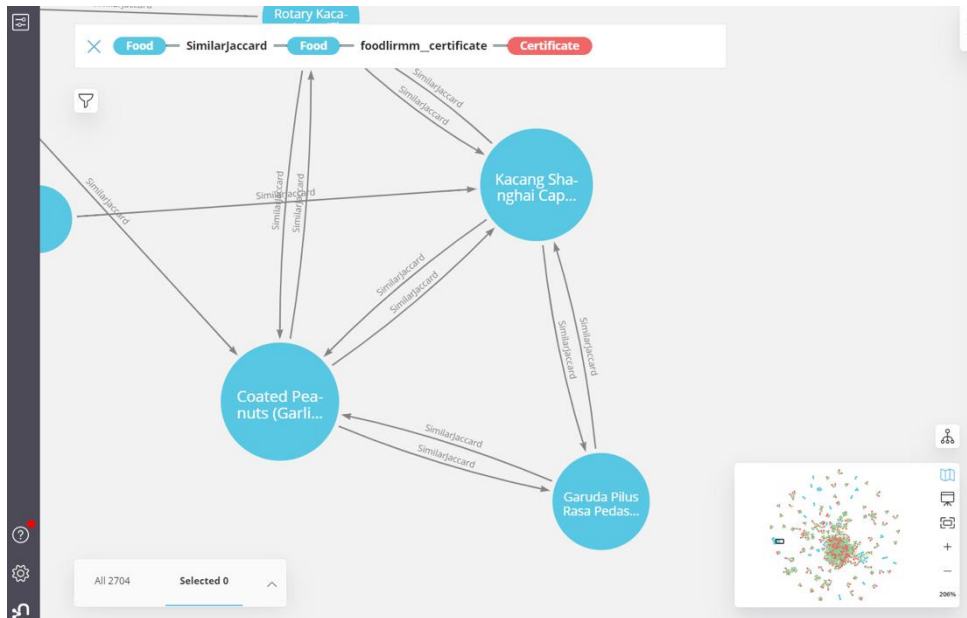
Figure 3. Visualization for Products without Halal Certificate (Jaccard Similarity)

When a product does not have a halal certificate but is related to a halal-certified product, it is considered halal and given a yellow color. As can be seen in Figure 4, *Tango Susu Vanilla* and *Wafer Selamat* are linked to *Dua Kelinci Deka Wafer Rolls*, which has a halal certificate.
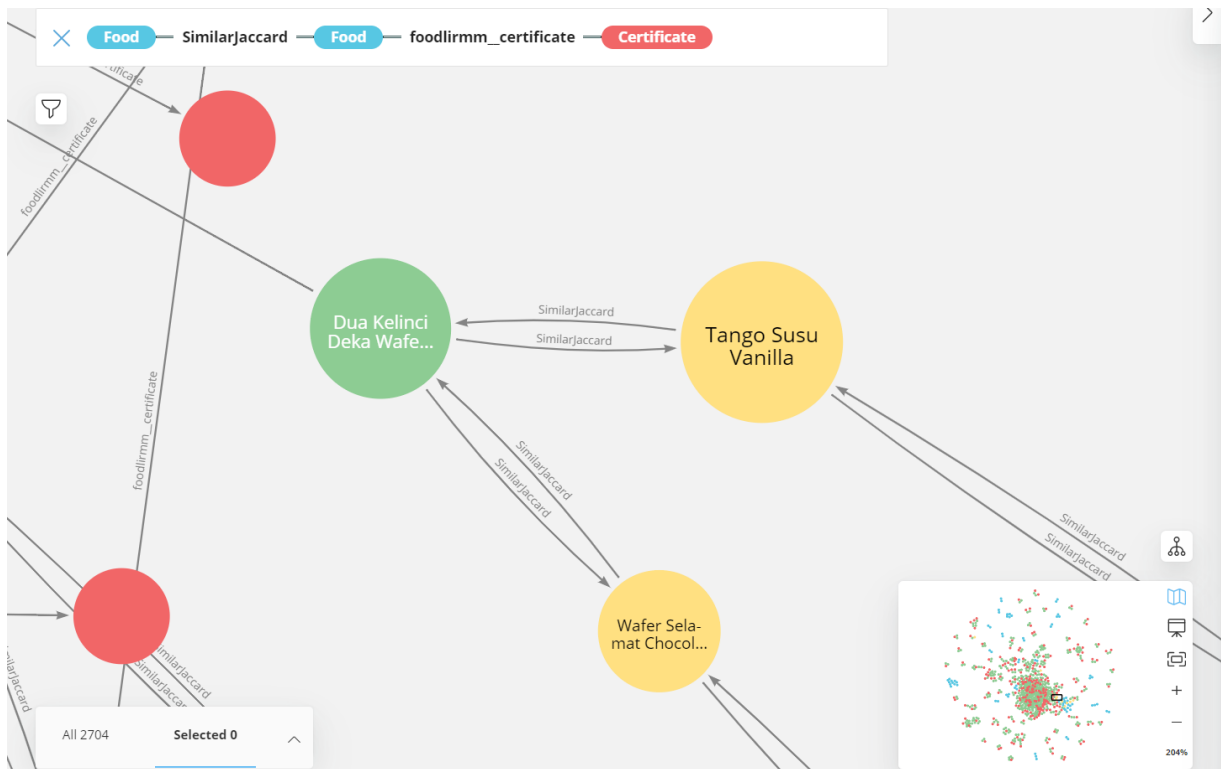

Figure 4. Two products connect to a halal-certified product.

4   Conclusions

We have presented our effort to link halal-certified products with non-halal certified products. Product data were taken from the KlikIndomaret and Halal MUI websites. Four algorithms were utilized to analyze the dataset, i.e., Jaccard Similarity, ANN, Adamic Adar, and Preferential Attachment. Three algorithms, i.e., Jaccard Similarity, ANN,

and Preferential Attachment, generated the same number of relationships between the products. Human evaluators assessed the algorithms' performance. The visualization can assist people, in particular Muslims, in predicting the halal status of food.

Based on our experience with the dataset from KlikIndomaret, a popular grocery retail chain in Indonesia, the number of non-halal-certified products is greater than the number of halal-certified products. Our results can recommend non-halal-certified products as alternatives for halal-certified products. The product analysis can be helpful for countries in judging the halal status of products where the traceability of products is not in place.

In addition, the results provide the basis for the development of a halal food database that can assist Muslims all over the world, even those who live in a non-Muslim country. Such a halal food database would not only contain food products that have been certified but also products that have not been officially certified. However, as the majority of world halal certification bodies only show the brand name, the manufacturer, and the expiry date of a product, there is a lack of data on the composition of halal food that is necessary for evaluating non-certified products (Aini Rakhmawati and Choirun Najib, 2020). It is, therefore, necessary to match product names and manufacturer names from halal certification institutions with data from open food databases such as OpenFoodFacts. According to data.world (data.world, 2022), 266 food datasets are available that could be assessed. Further problems may arise in situations where halal certificate institutions do not share halal-certified product names. As an example, Majelis Ulama Indonesia accredits 44 world halal certification bodies but only 22 percent have disclosed their data and only two institutions list food compositions (Kautsar and Rakhmawati, 2019).

We have used four algorithms to analyze the datasets. However, there are other similarity algorithms that could be used and may deliver different weights and values.

Finally, we give the following suggestions for halal certification institutions:

1. Halal certification institutions should share product names, company names and expiry dates, and the list of ingredients publicly.

2. The dataset should be created in a machine-readable format such as CSV. Most datasets are presented in HTML and PDF format, which need further pre-processing before using them.

3. A halal certification institution should provide an API for accessing the dataset in real time. Thus, we can analyze the data on the fly.

4. Several halal certification bodies provide their data in non-Latin characters such as in Arabic, Urdu, etc. They should show the dataset in two alphabets: Latin and non-Latin.

Acknowledgements

References

Adamic, L.A., Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, **25**(3): 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1.

Aini Rakhmawati, N., Choirun Najib, A. (2020). Discovering Entity Profiles Candidate for Entity Resolution on Linked Open Data Halal Food Products. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*. https://doi.org/10.1109/BigData50022.2020.9378411.

Albert, R., Barabási, A.L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**(1): 47. https://doi.org/10.1103/RevModPhys.74.47.

Benesty, J., Chen, J., Huang, Y., Cohen, I. (2009). Pearson Correlation Coefficient. *Springer Topics in Signal Processing*, **2**: 1–4. https://doi.org/10.1007/978-3-642-00296-0_5.

Çetin, G., Dinçer, M.Z. (2016). Muslim Friendly Tourism (MFT): A Discussion. *Journal of Tourismology*, **2**(1): 65–67. https://doi.org/10.26650/jot.2016.2.1.0005.

Data.world (2022). *Food data on data.world*. https://data.world/datasets/food.

Dong, W., Charikar, M., Li, K. (2011). Efficient K-nearest neighbor graph construction for generic similarity measures. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, 577–586. https://doi.org/10.1145/1963405.1963487

Grover, A., Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 855–864. https://doi.org/10.1145/2939672.2939754.

Gültekin, F., Çeker, O., Elgün, A., Gürbilek, M., Şimşek, M., Oral, R., Köse, S., Türker, S., Döndüren, H., Ünaldi, M., Küçüköner, E., Yetim, H., Aydin, A., Dede, B., Özmen, İ., Başayiğit, L., Cengiz, M., Tilki, T., Akin, S. (2020). Evaluation of Food Additives in Terms of Istihalah. *Akademik Platform Helal Yaşam Dergisi*, **2**(1): 1–13. https://dergipark.org.tr/tr/pub/apjhls/issue/55626/713197.

Hashemi, A., Dowlatshahi, M.B., Nezamabadi-pour, H. (2020). MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. *Expert Systems with Applications*, **142**: 113024. https://doi.org/10.1016/J.ESWA.2019.113024.

Ispirova, G., Eftimov, T., Seljak, B.K. (2020). P-nut: Predicting nutrient content from short text descriptions. *Mathematics*, **8**(10): 1811. https://doi.org/10.3390/math8101811.

Jia, X., Chaozhi, Z. (2021). Turning impediment into attraction: A supplier perspective on Halal food in non-Islamic destinations. *Journal of Destination Marketing & Management*, **19**: 100517. https://doi.org/10.1016/J.JDMM.2020.100517.

Karimah, N., Darwanto, D. (2021). Effect of Halal Awareness, Halal Certification, Food Ingredients Composition, Halal Marketing and Religiosity on Purchase Intention of Imported Food Products. *Al-Kharaj : Jurnal Ekonomi, Keuangan & Bisnis Syariah*, **3**(2): 177–191. https://doi.org/10.47467/alkharaj.v3i2.319.

Kautsar, A., Rakhmawati, N.A. (2019). *Integrasi Data Lembaga Sertifikasi Halal di Dunia Dengan Data Halal Nutrition Food Integration of Halal Certification Bodies Data from Around the World with Halal Nutrition Food Data*. Institut Teknologi Sepuluh November.

Kraemer, H.C. (2015). Kappa Coefficient. *Wiley StatsRef: Statistics Reference Online*, 1–4. https://doi.org/10.1002/9781118445112.STAT00365.PUB2.

Liu, D., Chen, X., Peng, D. (2019). Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets. *International Journal of Intelligent Systems*, **34**(7): 1572–1587. https://doi.org/10.1002/INT.22108.

Ministry of National Development Planning, & Committee. (2018). *MASTERPLAN EKONOMI SYARIAHINDONESIA 2019-202*. https://knks.go.id/storage/upload/1573459280-Masterplan Eksyar_Preview.pdf.

Mukidi, M. (2020). Prosedur Pemberian Sertifikat Label Halal Terhadap Produk Makanan Di Restoran Hotel Syariah Untuk Mewujudkan Hak Kenyamanan Konsumen Muslim (Studi Di Kementerian Agama Provinsi Sumatera Utara). *Jurnal Hukum Kaidah: Media Komunikasi Dan Informasi Hukum Dan Masyarakat*, **19**(3): 397–416. https://doi.org/10.30743/jhk.v19i3.2674.

Petković, M., Popovski, G., Seljak, B.K., Kocev, D., Eftimov, T. (2021). DIETHUB: Dietary habits analysis through understanding the content of recipes. *Trends in Food Science and Technology*, **107**: 183–194. https://doi.org/10.1016/j.tifs.2020.10.017.

Petriella, Y. (2019). *LPPOM UI: Baru 668.615 Produk di Indonesia Tersertifikasi Halal*. Bisnis.Com. https://ekonomi.bisnis.com/read/20190116/12/879298/lppom-ui-baru-668.615-produk-di-indonesia-tersertifikasi-halal.

Popovski, G., Seljak, B.K., Eftimov, T. (2019). FoodBase corpus: A new resource of annotated food entities. *Database*, **2019**(1): 121. https://doi.org/10.1093/database/baz121.

Rakhmawati, N.A., Firmansyah, A.A., Effendi, P.M., Abdillah, R., Cahyono, T.A. (2018). Auto Halal detection products based on Euclidian distance and cosine similarity. *International Journal on Advanced Science, Engineering and Information Technology*, **8**(4–2): 1706–1711.

Rakhmawati, N.A., Fatawi, J., Najib, A.C., Firmansyah, A.A. (2021). Linked open data for halal food products. *Journal of King Saud University - Computer and Information Sciences*, **33**(6): 728–739. https://doi.org/10.1016/J.JKSUCI.2019.04.004

Rakhmawati, N.A., Jannah, M. (2021). Food Ingredients Similarity Based on Conceptual and Textual Similarity. *Halal Research Journal*, **1**(2): 87–95.

Rezai, G., Mohamed, Z., Shamsudin, M.N. (2012). Assessment of Consumers' confidence on Halal labelled manufactured food in Malaysia. *Pertanika Journal of Social Science and Humanities*, **20**(1): 33–42.

Rezaul Karim, M., Cochez, M., Jares, J.B., Uddin, M., Beyan, O., Decker, S. (2019). Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. https://doi.org/10.1145/3307339.3342161

Salman, F., Kamran S. (2011). An exploratory study for measuring consumers awareness and perceptions towards halal food in Pakistan. *Interdisciplinary Journal of Contemporary Research in Business*, **3**(2): 639–651.

State of the Global Islamic Economy. (2021). State of the Global Islamic Economy Report 2020/2021. In *Dubai International Financial Centre*.

Svinarky, I., Malau, P. (2020). Penerbitan Sertifikat Halal Berdasarkan Prespektif Hukum Di Indonesia. *Jurnal Cahaya Keadilan*, **8**(1): 71–85. https://doi.org/10.33884/jck.v8i1.1896

Tieman, M., Williams, G. (2019). Creative Destruction of Halal Certification (Bodies) By Blockchain Technology? *ICR Journal*, **10**(1): 127–131. https://doi.org/10.52282/ICR.V10I1.79

Warto, W., Samsuri, S. (2020). Sertifikasi Halal dan Implikasinya Bagi Bisnis Produk Halal di Indonesia. *Al Maal: Journal of Islamic Economics and Banking*, **2**(1): 98–112. https://doi.org/10.31000/almaal.v2i1.2803

Yunus, R., Arif, O., Afzal, H., Amjad, M.F., Abbas, H., Bokhari, H.N., Haider, S.T., Zafar, N., Nawaz, R. (2019). A Framework to Estimate the Nutritional Value of Food in Real Time Using Deep Learning Techniques. *IEEE Access*, **7**: 2643–2652. https://doi.org/10.1109/ACCESS.2018.2879117