Halal Research **1 (2)** (2021) 87-95

Food ingredients similarity based on conceptual and textual similarity

Nur Aini Rakhmawati*, Miftahul Jannah

Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

ABSTRACT

Open Food Facts provides a database of food products such as product names, compositions, and additives, where everyone can contribute to add the data or reuse the existing data. The open food facts data are dirty and needs to be processed before storing the data to our system. To reduce redundancy in food ingredients data, we measure the similarity of ingredient food using two similarities: the conceptual similarity and textual similarity. The conceptual similarity measures the similarity between the two datasets by its word meaning (synonym), while the textual similarity is based on fuzzy string matching, namely Levenshtein distance, Jaro-Winkler distance, and Jaccard distance. Based on our evaluation, the combination of similarity measurements using textual and Wordnet similarity (conceptual) was the most optimal similarity method in food ingredients.

Keywords: *Halal, Jaccard distance, Jaro-Winkler distance, Levenshtein distance, Open food facts, Wordnet.*

1 Introduction

Institute for Foods, Drugs, and Cosmetics Indonesian Council of Ulama, known as LPPOM MUI, is an authorized halal certification institut in Indonesia. On the LPPOM MUI website, we can search for halal products and certificate information. As of this paper written, there were 727,617 halal certified products in LPPOM MUI data from 2011 to 2018. The information provided on the LPPOM MUI website only includes the product name, halal certificate number, manufacturer, and certificate validity date. In this search system, there is no access to information about the composition and nutrition of the product [1].

Linked Open Data Halal Food (LODHalal, http://halal.addi.is.its.ac.id/) is a research that collects food products especially halal certified products and publishes the data in Linked Data format. Besides that, Halal Nutrition Food also provides nutritional information and composition of these products [2], which cannot be found on the LPPOM MUI halal product search site. LODHalal provides a user interface where user can easily access the information above. To enrich the LODhalal, we add Open Food Facts data. Open Food Facts is a site that provides an open database of food product information where everyone in the world can

* Corresponding author. Tel: 031-5999944.
Email address: nur.aini@is.its.ac.id

Received August 22, 2021; Revised August 31, 2021; Accepted September 22, 2021

contribute to add product data to it. This database is published as open-data, so it can be reused by anyone [3].

To reduce the redundancy of food ingredients data in the LODHalal and Open Food Facts, it is necessary to measure the similarity of food ingredients. Therefore, the same food ingredients written in different terms can be standardized into the same terms. There are two methods proposed in this paper, namely the conceptual similarity and textual similarity. The conceptual similarity is used to measure the distance of synonyms between two different words/terms, while textual similarity measures the distance of similarity of characters between the two words. We adopted the semantic similarity using path-length similarity on WordNet which is proposed by Leacock and Chodorow [4]. WordNet similarity is also used by Martin Warin of Stockholm Universitet [5] to compare several semantic similarity measurement methods to enrich an ontology called the Common Procurement Vocabulary (CPV) using measures for semantic similarity and WordNet. Hongzhe Liu [6] employs Wordnet for calculating the similarity between text and very short sentences without using an external literary corpus.

2 Similarity methods

2.1 WordNet similarity path

WordNet is an English lexical database developed by Princeton University. Nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), where each set express different concepts. Synsets are interconnected with lexical relations. WordNet is also publicly available for download. The structure of WordNet is very useful for computational linguistics and natural language processing [7].

WordNet::Similarity::path is a module for calculating semantic relationships between words by counting word nodes in the 'is-a' hierarchy of WordNet. The length of a path consists of nodes. The longer the path the more unrelated the two words/concepts. Therefore, the similarity between the two concepts is inversely proportional to the path length (distance), which can be defined as in Eq. (1) [5].

$$similarity = \frac{1}{distance} \tag{1}$$

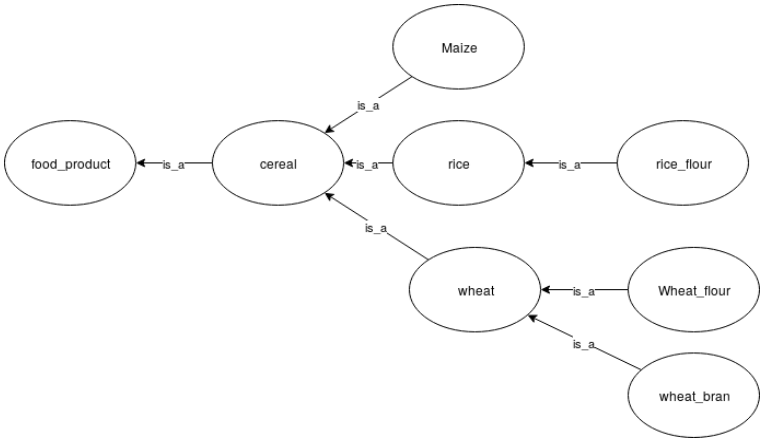

Fig. 1 Example of is-a hierarchy in Wordnet.

Figure 1 is an example illustration of the is-a hierarchy in WordNet. The distance is the shortest path between the two concepts. For instance, the distance between Wheat Flour and Wheat Bran is two. Therefore, the Wordnet similarity of Wheat Four and Wheat Bran is ½ = 0.5

2.2 Levenshtein Distance

Levenshtein distance uses edit distance in its operations. Edit distance calculates the minimum editing distance between strings to be compared with the target string. Levenshtein distance counts the number of additions, deletions, or character substitutions between two strings [8].

Example:

The Levenshtein distance between "kitten" and "sitting" is 3, because there are three edits to change one string to another, and there is no other way to do it in less than three edits:

1. kitten → sitten (substitution of character "s" against "k")

2. sitten → sittin (subtitution of character "i" against "e ")

3. sittin → sitting (addition of the "g" character at the end).

2.3 Jaro-Winkler Distance

The Jaro-Winkler algorithm is a similarity measuring algorithm between two strings. This algorithm is an algorithm discovered by Matthew A. Jaro in 1989 and 1995. This algorithm was later developed by William E. Winkler by modifying Jaro Distance to give a higher weight to the similarity prefix [9]. This algorithm performs:

1. Calculate the length of a string,
2. Count the number of same characters in two strings, and
3. Count the number of transpositions.

The Jaro-Winkler can be formulated in Equation 2 and 3.

$$d_j = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) & otherwise \end{cases} \quad (2)$$

where:

m is number of the same character.
|s1| is length of string 1.
|s2| is length of string 2.
t is number of transposition.

In the Jaro-Winkler (dw) algorithm, the prefix scale (p) is used to give the string set prefix. With the following formula:

$$d_w = d_j + (lp(1 - d_j)) \quad (3)$$

where:

dj is the result of calculating the similarity of the string s1 and s2 with the Jaro Distance algorithm.

l is the same character length at the beginning of the string before any inequality is found with a maximum limit of up to 4 characters.

P is the standard value for constants in Winkler's work p = 0.1.

2.4 Jaccard Distance

The Jaccard similarity coefficient is a statistical method used to compare the similarity and diversity of sample sets. The Jaccard coefficient measures the similarity between finite sample sets and is defined as the number of slices divided by the combined number of sample sets [10].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

where:

J (A,B) = $Jaccard$ Similarity Coefficient

A = Word A
B = Word B
Jaccard distance measures between sample sets is

dj (A,B) = 1- J (A,B) (5)

3 Methodology

After downloading the CSV dataset from the Open Food Facts, the steps of our methodology can be explained as follows:
1. Remove the duplicate products
2. Delete a product without ingredients information.
3. Transform the values to lowercase.
Table 1 is an example of a product with its ingredients.

Table 1. Example of food product after casefolding

Code	Product_name	Ingredients_text
1199	Solène céréales poulet	antioxydant: érythorbate de sodium, colorant: caramel - origine ue), tomate 33,3%, mayonnaise 11,1% (huile de colza 78,9%, eau, jaunes d'oeuf 6%, vinaigre, moutarde [eau, graines de moutarde, sel, vinaigre, curcuma], sel, dextrose, stabilisateur: gomme de cellulose, conservateur: sorbate de potassium, colorant: ? carotène, arôme)

4. Remove unnecessary characters such as punctuation (except comma) and percentage symbol (%). Table 2 shows the cleaned data after removing unnecessary characters.

Table 2. Example of food product after characters removal and replacement

Code	Ingredients text
1199	antioxydant, érythorbate de sodium, colorant, caramel, origine ue, tomate, mayonnaise, huile de colza, eau, jaunes d'oeuf, vinaigre, moutarde ,eau, graines de moutarde, sel, vinaigre, curcuma, sel, dextrose, stabilisateur, gomme de cellulose, conservateur, sorbate de potassium, colorant, carotene, aroma.

5. Translate non-English food ingredients into English

Some ingredients are found in French. The translation process uses the Google translate API. Table 3 is an example of food ingredients after the translation process.

Table 3. Example of food product after translating

Code	Ingredients_text
1199	antioxidant, sodium, sodium dyestuff, dye, caramel, eu origin, tomato, mayonnaise, rapeseed oil, water, egg yolks, vinegar, mustard, water, mustard seed, salt, vinegar, turmeric, salt, dextrose, stabilizer, cellulose gum, preservative, potassium sorbate, dye, carotene, aroma.

6. Removing stopword

There are three types of stopword namely ingredients, guidelines, and additional information. Some of ingredient stopwords are ingredient, contain, including, traces of, may contain, made from, that contain etc. Some of guidelines stopwords are to keep at, unopened, must be, to consume, preferably before, stored in, refrigerator, etc. Some of additional information stopwords are net weight, this is regularly, checked, during, minimum, durability, period, etc. Table 4 is an example of data after cleaning the stopwords.

Table 4. Example of food product after stopword removal

Code	Ingredients text
1199	antioxidant, sodium, sodium dyestuff, dye, caramel, eu origin, tomato, mayonnaise, rapeseed oil, water, egg yolks, vinegar, mustard, water, mustard seed, salt, vinegar, turmeric, salt, dextrose, stabilizer, cellulose gum, preservative, potassium sorbate, dye, carotene, aroma.

7. Tokenization

The tokenization process separates ingredients by comma (,). An example of data after the tokenization process is shown in Table 5.

Table 5. Example of list of ingredients after tokenization

Ingredients
Antioxidant
Sodium
Sodium dyestuff
Dye
Caramel
Eu origin
Tomato
Mayonnaise
Rapeseed oil
Water

8. Ingredients similarity

After pre-processing of food ingredients, the similarity of food ingredients is measured using two types of similarity measurement methods, namely conceptual and textual. Conceptual measurement uses Wordnet with Leacock Chodorow (LCH) method. Textual similarity measurements use the method of Levenshtein distance, Jaro-Winkler distance, and Jaccard distance. It will also combine two conceptual-textual methods, namely Wordnet & Jaccard, Wordnet & Jaro-Winkler, and Wordnet & Levenshtein. The similarity is used for comparing ingredients between open food facts data and LODHalal data. As seen in Figure 4, after calculating the similarity, only ingredients that passes the similarity threshold 80% store to a temporary database.

4 Evaluation

We evaluate the similarity based on precision and recall. Recall is the level of success of the system in providing correct information, while precision is the level of accuracy between the actual information and the results provided by the system. We combine the value of precision and recall in F Measure metric (Eq. 6)

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

For Wordnet LCH, we validate the result based on the Dictionary of Food Ingredients Fourth Edition [11] and Google search.

5 Results and discussion

271,823 food ingredients are measured the similarity value between of them. As seen in Table 4, Jaro Winkler generates set of pair ingredients with 80% similarity value which is higher than other methods. The textual similarity performs better than conceptual similarity. This is due to several factors, such as Wordnet LCH does not process words that are not contained in the Wordnet lexical database. In addition, the miss-typed food words (example: ‘pinrapple’; which should be ‘pineapple’) will not enter the process of measuring the Wordnet LCH similarity. Therefore, the combination method textual similarity and conceptual similarity increase the number of pair similarity data. The F-measure of all methods ranges from 50%

to 65%. The recall of the Wordnet LCH is higher than other methods since the Wordnet LCH has a hierarchy of words that can interpret words meaningfully not textually. The Levenshtein distance has succeeded in providing a high degree of accuracy in the similarity of food that is supposed to be exactly the same.

Table 6. All results of precision and recall

Method	Number of data pairing	Precision	Recall	F-Measure
Jacard Distance	757,440	80.6%	44.6%	0.574%
Jaro-Winkler Distance	31,271,070	82.9%	51.8%	0.637%
Levenshtein Distance	57,044	100.0%	46.4%	0.633%
Wordnet LCH similarity	3,085	41.6%	75.0%	0.535%
Wordnet & Jaccard	759,976	40.9%	80.4%	0.542%
Wordnet & Jaro-Winkler	31,273,531	40.9%	80.4%	0.542%
Wordnet & Levenshtein	60,049	41.6%	75.0%	0.535%

According to Figure 3, 4, and 5, the median of distribution of pair data for textual similarity is between 0.825 and 0.850. Many pairs of data for Jaro-Winkler and Jaccard are in outlier area but above 90%. The distribution of pair data of Wordnet LCH spread across from 80% to 100%.

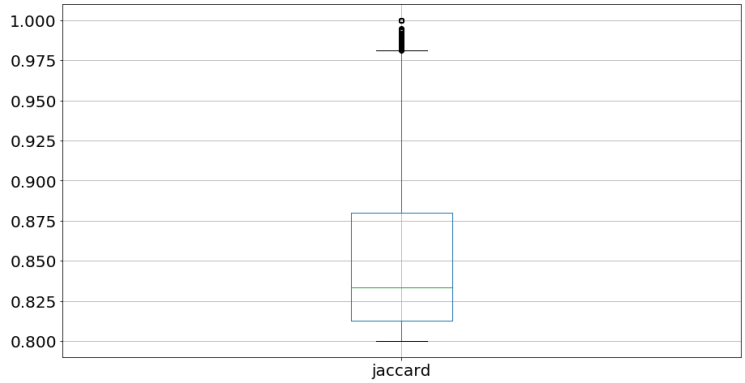

Fig. 2 Distribution of pair data for Jaccard

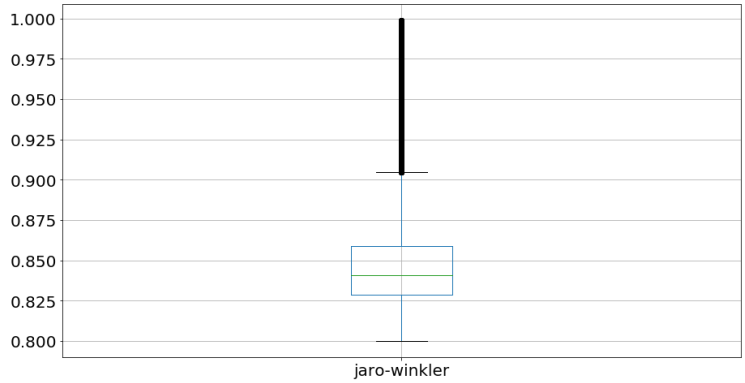

Fig. 3 Distribution of pair data for Jaro-Winkler

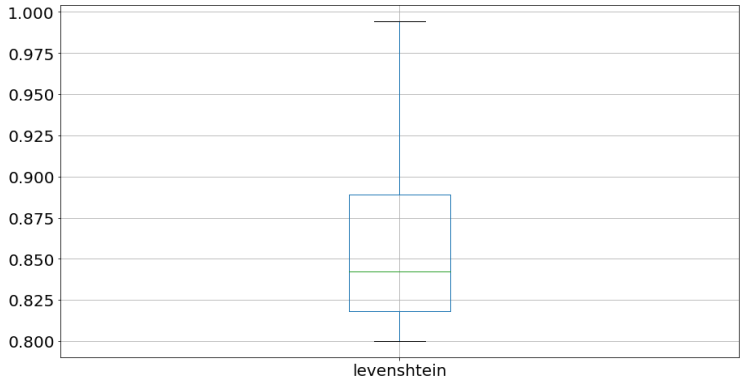

Fig. 4 Distribution of pair data for Levenshtein

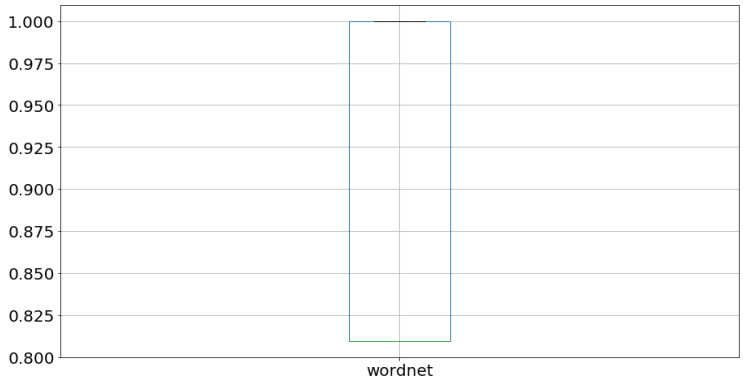

Fig. 5 Distribution of pair data for Wordnet LCH

After the implementation of the similarity measurement, the number of ingredients was decreased to 83.531 from 271,823.

6 Conclusion

We have presented how we measure the similarity of food ingredients in Open Food Facts and LODHalal by using textual and conceptual similarity. By using these methods, we can reduce the number of redundancy data about 70%. Levenshtein distance is the most accurate method compared to Jaro-Winkler and Jaccard distance. This is also supported by the complexity of Levenshtein algorithm compared to the two textual measurement methods. Levenshtein considers in editing distance insertion, deletion, and substitution in the measurement algorithm. The combination textual and conceptual can increase the recall value of textual similarity of words and word meanings. In the future, we can use other methods, such as Guessoum. Furthermore, the search feature also applies similarity in the LODHalal application.

References

[1] LPPOM MUI. http://www.halalmui.org/mui14/index.php/main/go_to_section/130/1511/page/1. [Accessed: February 1, 2019].

[2] N.A. Rakhmawati, J. Fatawi, A.C. Najib, and A.A. Firmansyah. "Linked open data for halal food products," *J. King Saud Univ. - Comput. Inf. Sci.* 2019, vol. 33, no. 6, pp. 728-739.
[3] Open Food Facts. https://world.openfoodfacts.org/discover [Accessed: January 18, 2019]
[4] C. Leacock, M. Chodorow, G.A. Miller, "Combining local context and WordNet similarity for word sense identification," In *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed., Cambridge, MA, USA: MIT Press, 1998, pp. 265-283.
[5] M. Warin. Using WordNet and Semantic Similarity to Disambiguate an Ontology, 2004.
[6] H. Liu and P. Wang, "Assessing Sentence Similarity Using WordNet based Word Similarity," *J. Softw*. 2013, vol. 6, no. 6, pp. 1451-1458.
[7] WordNet (2019): a lexical database for English. (n.d.). https://wordnet.princeton.edu/ [Accessed: January 18, 2019]
[8] C. Zhao and S. Sahni, "String correction using the Damerau-Levenshtein distance," *BMC Bioinform*. 2019, vol. 20, 277.
[9] F. Friendly, "Jaro-Winkler distance improvement for approximate string search using indexing data for multiuser application," *J. Phys. Conf. Ser.* 2019, vol. 1361, no. 1.
[10] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull. Soc. Vaudoise Sci. Nat*. 1901, vol. 37, pp. 547-579.
[11] R.S. Igoe and Y.H. Hui, *Dictionary of Food Ingredients Fourth Edition*. Gaithersburg: Aspen Publishers, Inc., 2001.
[12] D. Guessoum, M. Miraoui, and C. Tadj, "A modification of wu and palmer semantic similarity measure," Conf. UBICOMM 2016, The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2016, pp. 42-46.