p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

Query Suggestion on Drugs e-Dictionary Using the Levenshtein Distance Algorithm

Halimah Tus Sadiah[a1], Muhamad Saad Nurul Ishlah[a2], Nisa Najwa Rokhmah[b3]

[a]Manajemen Informatika, Universitas Pakuan
Jl.Pakuan, Bogor 16143
[1]sadiahht@unpak.ac.id (Corresponding author)
[2]nurul.ishlah@unpak.a.c.id

[b]Farmasi, Universitas Pakuan
Jl.Pakuan, Bogor 16143
[3]nisanajwarokhmah@gmail.com

Abstract

Dictionary of medicine in the form of a thick book has many disadvantages, one of which is impractical. This is the reason for Indonesian developers to create drugs e-Dictionary. But the drugs e-Dictionary that has been developed is still in the form of a letter index so that users must search the terms one by one in sequential order. This has become so inefficient and ineffective that it is necessary to add a search function and query suggestion feature to the drug e-dictionary. The purpose of this study is to build a query suggestion facility on drugs e-Dictionary using the Levenshtein Distance algorithm. The stages of this research consist of the Development of web-based drugs e-Dictionary, Implementation of the Levenshtein Distance Algorithm, Query Suggestion Testing, and Usage. The query suggestion function works by producing the closest word output contained in the database. Based on the results of the implementation of the Levenshtein Distance algorithm and test results, Drugs e-Dictionary can evaluate words that are not in the database. It reaches 90% accuracy of inputted query, with 90% precision and 90% recall in confusion matrix.

Keywords: *Query Suggestion, Drugs e-Dictionary, Algorithm, Levenshtein Distance Algorithm*

1. Introduction

The decision to use a drug (medication drug) always raise concern on the benefits and its risks so that a Pharmacist needs a drug dictionary to search for previously unknown terms of medicine [1]. Besides, the drug dictionary becomes one of the learning tools that are used by Pharmacists, Students and Indonesian community in learning medicine or foreign terms about medicine. The drug dictionary that is used nowadays is in the form of a thick physical dictionary book. It turns out to have drawbacks, such as it is too heavy to be carried so that it is not practically handy. This is one of many reasons for Indonesian developers to compete in creating an electronic dictionary of drugs or what we know as the term drug e-dictionary.

Most of the available drugs e-dictionaries that have been developed so far are still in the form of a letter-index based dictionary, it makes users have to search for words or terms one by one in a sequential fashion. This has become so inefficient and ineffective that it is necessary to add a search function to the drug e-dictionary. The search function on drugs e-dictionary is very important because it can be used as a shortcut when searching words or terms needed so that users can search for words effectively and efficiently [2][3].

The drug e-dictionary search function needs to be optimized with the addition of the Query Suggestion facility. Query Suggestion is some kind of an interface between a user and a search engine [4]. This facility is an effective and efficient approach to help the user in the process of finding information by providing a suggestion for the user when mistyping is happened in the search form [2][3][5][6]. This feature is very important to be applied since it can improve the usability factor of searching [7][8].It works by looking for the similarity between a correct query and false query in a database [9].This feature can be a solution for preventing the user from typing

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

the wrong name of the drug. The Query Suggestion can be used in a search application by implementing the Levenshtein Distance Algorithm.

The Levenshtein Distance algorithm is an algorithm created by Vladimir Levenshtein in 1965 [10]. This algorithm looks for the distance between the words entered by the user and the words stored in the system database by the method of calculating the number of differences between the two strings in the form of a matrix [11][12]. It works by calculating the distance between the two strings and then look for the minimum number of change operations to change from string A to string B. The calculation is represented using the Levenshtein distance calculation table, where the last value in the lower right corner is the final value of the second distance string. In the Levenshtein distance algorithm there are three operations performed, namely the operation of changing characters, adding characters and deleting characters [13[14] [15].

Research on query suggestion has been done by Jiang et al (2008), namely Query suggestion by query search: a new approach to user support in web search [3]. Meanwhile, research on the Levenshtein Distance algorithm was conducted by Ngafidin and Wibawanto (2015), namely the Implementation of the Autocomplete Feature and the Levenshtein Distance Algorithm to Increase the Effectiveness of Word Search in the Indonesian Big Dictionary (KBBI) [16]. This study aims to build a query suggestion facility using the Levenshtein Distance algorithm on drugs e-dictionary. This research is very important to do so that pharmacists, students, and the public can easily search for drug terms in the drug e-dictionary.

2. Reseach Methods

The research method used in this study consists of several stages, as shown in figure 1, which is described below:

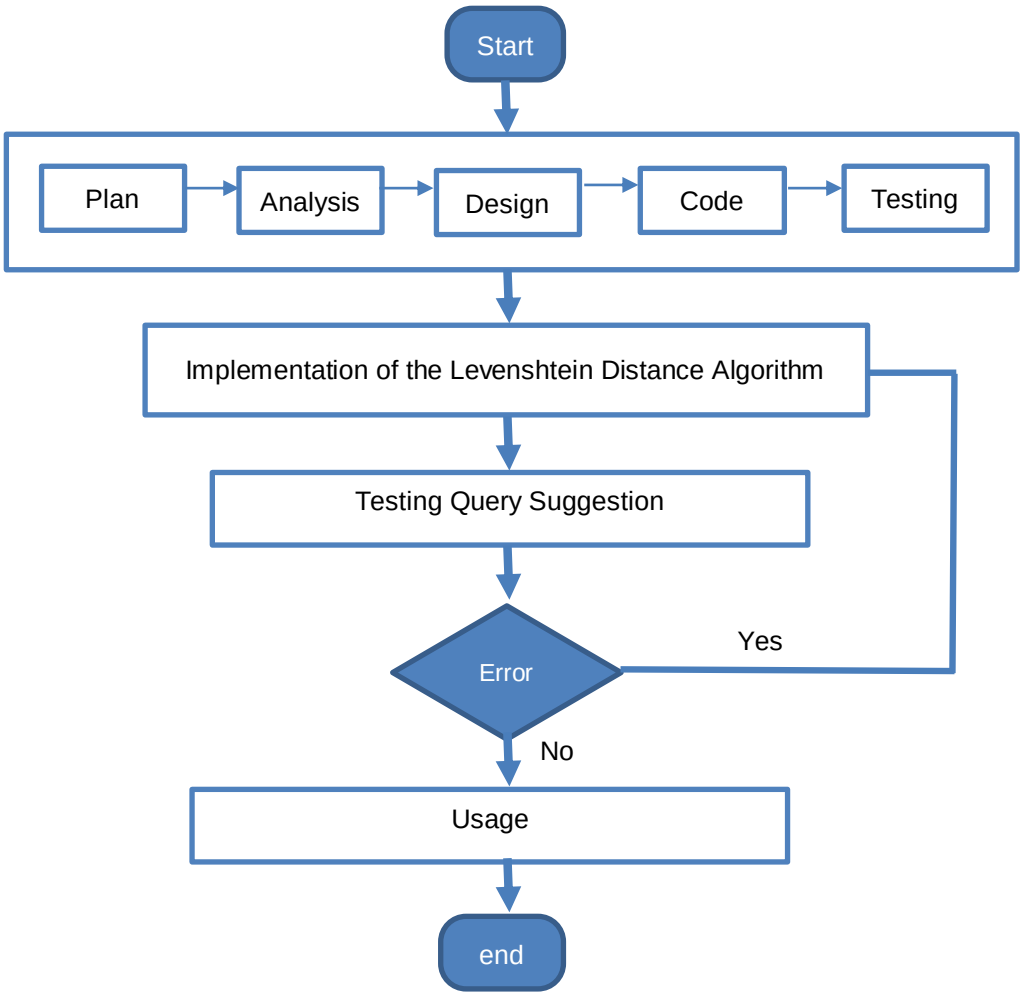

Figure 1. Research Method

1. Development of Web-based e-Dictionary Drugs
   Development of Web-based e-Dictionary Drugs using the SDLC (System Development Life Cycle) method that has been adapted to the needs of Web-based Drugs e Dictionary [17][18]. The stages are Plan, Analysis, Design, Code, Testing.
2. Implementation of the Levenshtein Distance Algorithm
   The implementation is done by adding the Levenshtein Distance algorithm in PHP programming language.
3. Testing Query Suggestion
   Testing is done by inputting drug terms in the search form as many as 100 terms. The number of terms entered consists of 50 correct terms, 50 incorrect terms, or incorrect terms.
4. Usage
   Drugs e-Dictionary that has been tested are then hosted to be used by users.

3. Result and Discussion

3.1. Web-based Drugs e-Dictionary Development

3.1.1. Plan

In the planning stage, data collection is carried out. Data was collected from the ISO Indonesian Information Specialist book [19]. The collected data consist of drug categories, drug names, indications, contradictions, side effects, drug interactions, dosages, packaging, and drugs warning.

3.1.2. Analysis

In this stage, system functionality requirements and non-system requirements are collected. There are 28 system functionality requirements, namely 10 front end system functionality requirements and 18 back end system functionality requirements. The non-functional requirements only produced 7 system non-functional requirements.

3.1.3. Design

Next in this design stage, a search system flow will be developed. It is depicted in figure 2. Based on figure 2, the flow of drug searching is described below:

1. User accesses drugs e-dictionary website
2. User searches for name or drug term in the search form
   - If user's query is empty, then system will show empty query notification or "query has not been inserted".
   - If inputted query is in the database, then system will show search result.
   - If inputted query is not available in the database, then system will proceed with Levenshtein Distance Algorithm followed up with the query suggestion

3.1.4. Code

In this implementation stage, the system is developed in PHP language with MYSQLi for database connection. The result is the web *drugs e-dictionary. Drugs e-dictionary* consists of main searching page, which search based on drugs term as depicted in figure 3; searching based on disease indication, as shown in Figure 4; and A-Z index-based searching, as depicted in Figure 5.

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

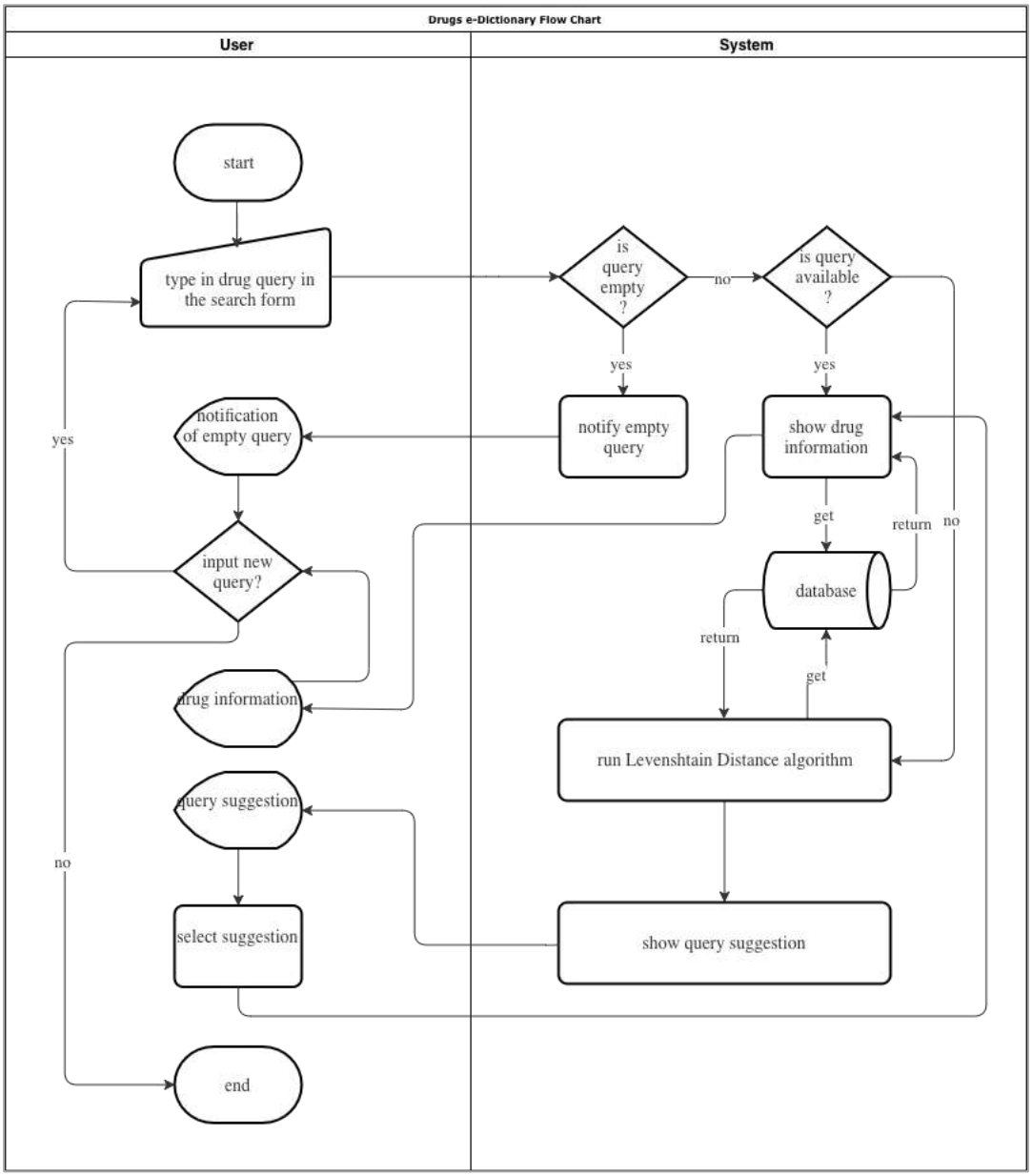

Figure 2. Developed Search System Flow

Figure 3. Drugs e-Dictionary website

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

Figure 4. Homepage user interface based disease indication

Figure 5. A-Z index-based searching page

3.1.5. Testing

Black box is used in the testing stage. It is usually called as system functional test [11]. Based on the test, 28 functions from the system is running as expected.

3.2 Levenshtein Distance Algorithm Implementation

Figure 6 is the pseudocode of the Levenshtein distance algorithm.

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

```
int LevenshteinDistance(char s[1..m], char t[1..n])
{
// d is a table with m+1 rows and n+1 columns
declare int d[0..m, 0..n]
declare int cost
for i from 0 to m
d[i, 0] := i // deletion
for j from 0 to n
d[0, j] := j // insertion
for j from 1 to n
{
for i from 1 to m
{
if s[i] ≠ t[j] then
cost := 1
else
cost := 0
d[i, j] := minimum
(
d[i-1, j] + 1, // deletion
d[i, j-1] + 1, // insertion
d[i-1, j-1] + cost // substitution
)
}
}
return d[m,n]
```

Figure 6. Pseudocode of the Levenshtein Distance Algorithm

Pseudocode of the algorithm, as depicted in figure 6, can be computed manually as shown in figure 7 and figure 8.

Let “Paraci” be inputted characters, and word in the database is Paraco

M = Inputted by user = Paraci
N = Word in the database = Paraco
D [0,0] = 0

Initialize first row and first column with 0,1,2,. . . m 0,1,2,...n

		P	A	R	A	C	O
	0	1	2	3	4	5	6
P	1	0					
A	2	1					
R	3	2					
A	4	3					
C	5	4					
I	6	5					

Figure 7. First row and column initialization

- For each character, compare each character from inputted word with actual word in database. If it is a match, then the cost is 0, otherwise the cost will be 1
- Check the minimum,

	Top	=	d[i,j]+1
d[i,j]	Side	=	d[i,j]+1
	Diagonal	=	d[i,j]+ cost

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

- Compare character P with P, put cost = 1 if differ, otherwise cost =0

  Check all values in d [i,j] — Top = 1, Diagonal = 0, Side = — Minimum diagonal

- d [i,j] = d [i,j] + cost = 0 + 0 = 0 so on, so forth

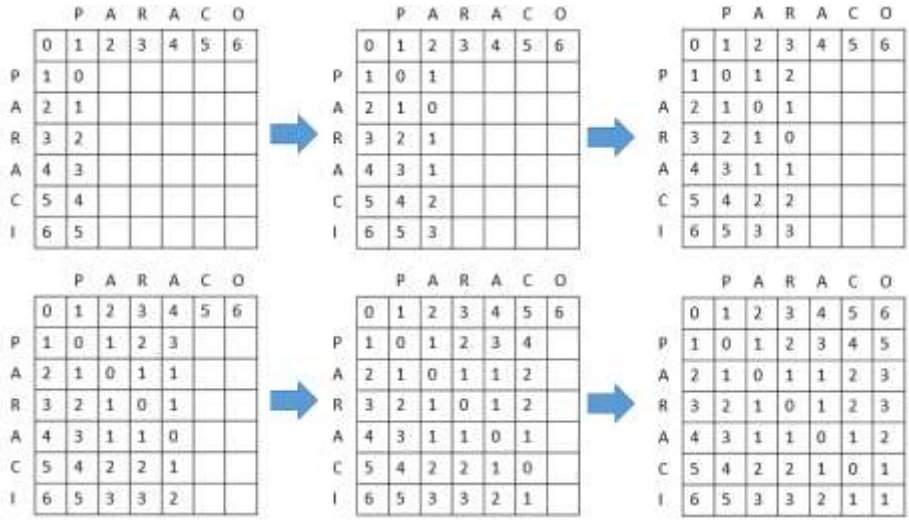

Figure 8. Manual computation process of *Levenshtein Distance* algorithm

In the Figure 8, the distance generated is a value that is in the lower-right corner of the matrix, which is 1. The value of one means there is 1 operation performed. The value of one is generated from the operation of the sum of the cost values with a diagonal minimum value. The distance value obtained from the diagonal side means that the operation that works is a substitution. So for the PARACI String to be converted into a PARACO string, one operation is needed, namely the substitution of the first character ("I") to the character "O" so that the value of the Levenshtein Distance is equal to 1. The result of Levenshtein Distance algorithm implementation on drugs e-dictionary, as depicted in figure 9.

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

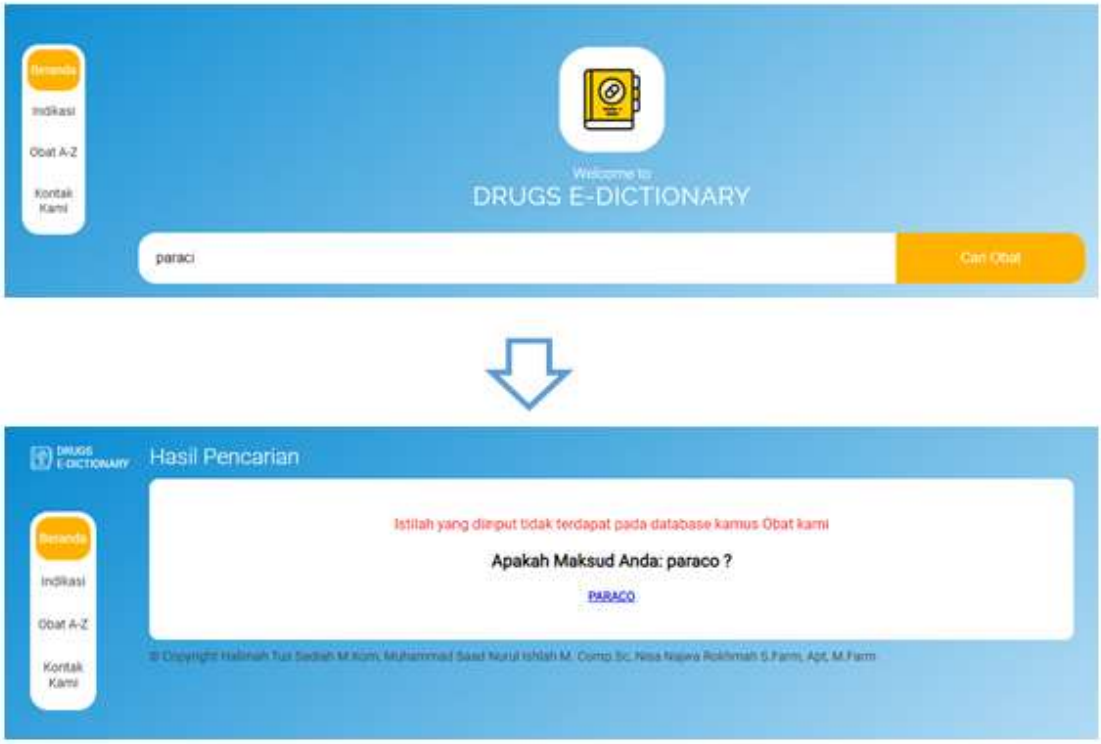

Figure 9. Result of Levenshtein Distance algorithm implementation on drugs e-dictionary

Testing the Drugs e-Dictionary with Query Suggestion Added Facility

Tests carried out in the form of validation testing by inputting 100 test queries into the search form. Table 1 summarizes the results of the validation test on drugs e-dictionary. Figure 9 shows an example of query suggestion testing.

Table 1. Example of Words in query suggestion validation testing

No	Inputted Query (drug name)	*Query Suggestion*	*Output*	Category of Levenshtein Distance Algorithm Operation	Notes	Validation
1	Paracetamol	-	Paracetamol	-	Inputted query is correct	Valid
2	Kamols	*"Apakah maksud anda kamolas, ?"*	Kamolas	Add letter a	Incorrect query inputted, lacking letters	Valid
3	Zephanall	*"Apakah maksud anda Zephanal?"*	Zephanal	delete letter l	Incorrect query entered, Excess letters	Valid
4	Paraci	*"Apakah maksud*	Paraco	Substitute i with o	Incorrect query entered	Valid

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

No	Inputted Query (drug name)	*Query Suggestion*	*Output*	Category of Levenshtein Distance Algorithm Operation	Notes	Validation
5	Diparin	*anda paraco?"* *"Apakah maksud anda Dapyrin ?"*	Dapyrin	*Query suggestion* by closest word	The inputted query does not exist in the database	Valid

In the Table 1, validation tests are categorized into insert, delete, and substitution operations. Whenever inputted query is not in database, the system will show notification of "The inputted query does not exist in the database" which then will show *Query suggestion* by generating some terms that are closer in the database. Let's take "diparin" as inputted query (unknown term in database), the system will show "dapyrin" as the suggestion (Table 1).

The developed system uses a non-case sensitive query checking. Hence it will not affect the output, whether the inputted query is in an uppercase or lowercase. In addition, if inputted query is a meaningless word, such as "ZZZZ", then the system will show a word with initial letter Z that has the fewest number of words in database, in this case "Zalona". The system will search for any terms with minimal Levenshtein Disctance algorithm operation.

To evaluate the accuracy of the searched query, it is represented in a confusion matrix (Table 2), which has four classification process results, namely: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN) [20].

Table 2. Confusion Matrix

Total Population	Predicted: yes	Predicted: no
Actual: yes	TP	FN
Actual: no	FP	TN

Based on TN, FP, FN and TP accuracy is obtained (equation 1), precision (equation 2) and *recall* (equation 3). Based on equation 1, equation 2 and equation 3, we have a result of query accuracy for drug query of 90%, precision=90%, recall = 90%. Confusion matrix for evaluation result of drug terms is in Table 3.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} * 100\% \quad 1)$$

$$precision = \frac{TP}{TP + FP} * 100\% \quad 2)$$

$$recall = \frac{TP}{TP + FN} * 100\% \quad 3)$$

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

Table 3. Confusion Matrix for Drug Terms Evaluation

100	Predicted: yes	Predicted: no
Actual: yes	45	5
Actual: no	5	45

In this research, a comparable test was conducted between the Levenshtein algorithm and Soundex to see which algorithm could give suggestion better for several test categories: addition, deletion, and letters substitution (Table 4). The result shows that the Levenshtein algorithm is able to give improvement suggestion for those three operations, while the Soundex could not give accurate suggestion for substitution, addition, and deletion of the first two letters in a word. However, Soundex algorithm can provide suggestions on words whose pronunciation is almost similar [21][22]. Figure 10 show the precision and recall comparation for both algorithms.

Table 4. Levenshtein Distance Algorithm and Soundex tests for several words

No	Inputted query	Expected outpur	Levenshtein Distance Algorithm		Soundex			Test category
					Inputted query code	Output code	Notes	
1	Kamols	Kamolas	Kamolas	valid	K542	K542	valid	Add 1 letter
2	Zephanall	Zephanal	Zephanal	valid	Z154	Z154	valid	Remove 1 letter
3	Paraci	Paraco	Paraco	valid	P620	P620	valid	Subtitute 1 letter
4	Paracetam	Paracetamol	Paracetamol	valid	P623	P623	Valid	Remove the last 2 letters
5	Lopurinol	allopurinol	Allopurinol	valid	L165	A416	Tidak Valid	Remove the first 2 letters
6	Afibram	afibramol	Afibramol	Valid	A116	A116	Valid	Add 2 last letters
7	Zoldine	zacoldine	Zacoldine	valid	Z435	Z243	Tidak valid	Add 2 first letters
8	Dilona	zelona	Zelona	valid	D450	Z450	Tidak Valid	Subtitute 2 first letters
9	Bimasil	Bimacyl	Bimacyl	Valid	B524	B524	valid	Subtitute 2 last letters
10	Unicetabil	unnicetamol	Unicetamol	valid	U523	U523	Valid	Subtitute 2 last letters

p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

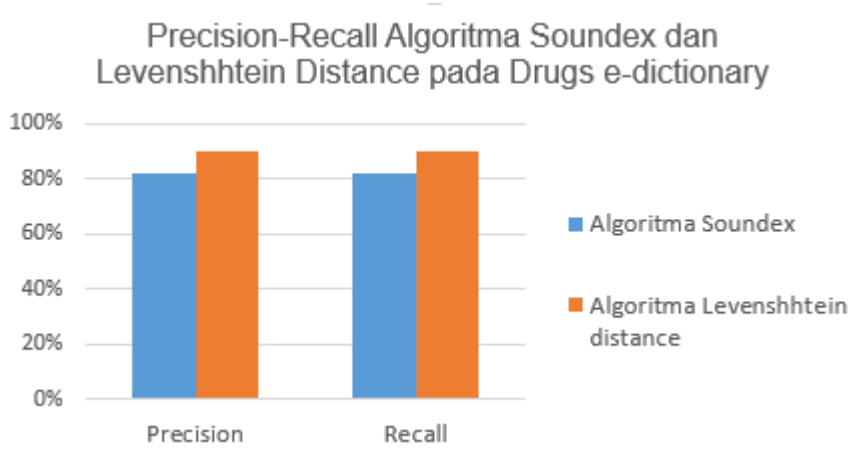

Figure 10 Precision-Recall Soundex and Levenshtein Distance algorithm in drugs e-dictionary application

Based on the Levenshtein distance algorithm test result, we conducted further comparison with previous studies as shown in Table 5. Based on the table 5, there is an improvement in precision for the Levenshtein algorithm.

Table 5. Comparison of Levenshtein distance algorithm performance with previous studies

No	Studies (year)	Soundex Algorithm		Levenshtein Distance Algorithm	
		Precision	Recall	Precision	Recall
1	*Koreksi Ejaan Istilah Komputer* (Pahdi 2016)	74	-	70	-
2	*Sistem Pencarian Ayat Al-Quran* (Arsaningtyas *et al*. 2018)	82	80	79	91
3	*Querry Suggestion pada drugs e-dictionary* (Sadiah *et al* 2019)	82	82	90	90

4. Conclusion

The drug e-dictionary search function needs to be optimized with the addition of the Query Suggestion facility. The Query Suggestion facility was developed using the Levenshtein Distance algorithm. Based on the results of the implementation, the Levenshtein Distance algorithm runs from the top left corner of a two-dimensional array that has been filled with several initial string characters and target strings and is given a cost value. The cost value at the lower right-hand end is the Distance edit value that represents the number of operations that the algorithm has to process. Based on the test results, the system can evaluate words that are not in the database with the query suggestion function closest to the database. It reaches 90% accuracy of inputted query, with 90% precision and 90% recall in confusion matrix. The next research is the implementation of the n gram algorithm on drugs e-dictionary and conducting comparative analysis research between the n-gram algorithm with the levenshtein distance algorithm.

Acknowledgment

This research was supported by KEMENRISTEKDIKTI, and LLDIKTI who provided the funding of this research. The author would also like to show gratitude to LPPM Universitas Pakuan for helping and facilitating this research.

References

[1] Departemen Kesehatan RI. Tanggung Jawab Apoteker Terhadap Keselamatan Pasien (Patient Safety). Jakarta: Direktorat Bina Farmasi Komunitas Dan Klinik Ditjen Bina Kefarmasian Dan Alat Kesehatan Departemen Kesehatan RI. 2008.

[2] Y. Song, & Li-wei He. 2010. Optimal Rare Query Suggestion With Implicit User. ACM Journals.pp: 901-910.
[3] S. Jiang, S. Zilles, & R. Holte. 2008. Query suggestion by query search: a new approach to user support in web search [Online]. [Cited 2018 August 1]. Available from www.cs.uregina.ca/~zilles/jiangZH09.pdf
[4] Y. Song, D. Zhou., & L.W. He. 2011. Post-ranking-query-suggestion-by-diversifying-searchresul [Online]. [Cited 2018 August 1]. Available from https://www.microsoft.com/id-id/: https://www.microsoft.com/en-us/research/publication/post-ranking-query-suggestionbydiversifying-search-results/
[5] J.-M.Yangy, R. Cai, F. Jingz, S.Wangy, L. Zhangy, & W.Y.Ma. 2008. Search-based Query Suggestion.[Online] [Cited 2018 August 1]. Available from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.3499&rep=rep1&type=pdf
[6] Q. Mei, D. Zhou & K Church 2008. Query Suggestion Using Hitting Time [Online]. [Cited 2018 August 1]. Available from https://www.microsoft.com/enus/ research/wpcontent/uploads/2017/01/sugg.pdf
[7] H. Cao, D. Jiang,J. Pei, Q. He, Z. Liao, E. Chen, & H. Li. 2008. Context-Aware Query Suggestion by Mining Click-Through [Online]. [Cited 2018 August 1]. Available from https://www.cs.sfu.ca/~jpei/publications/QuerySuggestion-KDD08.pdf
[8] Z.-J. Zha, L. Yang, T. Me., M. Wang, & Zengfu. Visual Query Suggestion. *ACM Journals*. pp. 15-24. 2009.
[9] S. Bathia, D. Majumdar, & P. Mitra. Query Suggestions in the Absence of Query Logs. *ACM Journals*, pp. 1-10.2011.
[10] Z.Afriansyah, D.Puspitaningrum, & Ernawati. Rancang Bangun Aplikasi Pencocokan DNA Manusia Menggunakan Algoritma Levenshtein Distance (Studi Kasus: Dna Kanker Hati Manusia). *Jurnal Rekursif* . 3:2,pp. 61-67.2015.
[11] B. Pratama & S. Pamungkas, Analisis Kinerja Algoritma Levenshtein Distance Dalam Mendeteksi Kemiripan Dokumen Teks. *Jurnal Log!k@* . 6:2, pp. 131-143.2016
[12] T. Aprilianto, & A. Badawi. Sistem Koreksi Kata Dan Pengenalan Struktur Kalimat Berbahasa Indonesia Dengan Pendekatan Kamus Berbasis Levenshtein Distance. *Jurnal SPIRIT*. 9:1, pp 48-61. 2017.
[13] R. Haldar, & D. Mukhopadhyay. 2011. Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach [Online]. [Cited 2018 August 1]. Available from
[14] R. Mishra, & N. Kaur. A Survey of Spelling Error Detection and Correction Techniques. *International Journal of Computer Trends and Technology*. 3:4, pp. 372-374. 2013
[15] N. Ariyani, N., R. Sutardi, & Ramadhan. Aplikasi Pendeteksi Kemiripan Isi Teks Dokumen Menggunakan Metode Levenshtein Distance. *semanTIK*. 2:1,pp. 279-286. 2016.
[16] K.N. Ngafidin & H. Wibawanto. Implementasi Fitur Autocomplete dan Algoritma Levenshtein Distance untuk Meningkatkan Efektivitas Pencarian Kata di Kamus Besar Bahasa Indonesia (KBBI). *Jurnal Teknik Elektro*. 7:1, pp.1-6. 2015
[17] R Pressman dan B.R. Maxim. Software Engineering a Practitioners approach. McGraw-Hill Education : New York. 2014.
[18] J Satzinger, R. Jackson, & S. Burd. System Analysis and Design in a changing World. USA: Course Technology Cengage Learning. 2010.
[19] Ikatan Apoteker Indonesia. ISO Informasi Spesialite Obat Indonesia Vol 52-Tahun 2019. Jakarta : Isfi Penerbitan.2019
[20] M. Navin, Pankaja R. Performance Analysis of Text Classification Algorithms using Confusion Matrix. *International Journal of Engineering and Technical Research (IJETR)*. 6:2,pp. 75-78. 2016
[21] A.Pahdi. Koreksi Ejaan Istilah Komputer Berbasis Kombinasi Algoritma Damerau-Levenshtein dan Algoritma Soundex. *Journal Speed – Sentra Penelitian Engineering dan* Edukasi.8:2, pp. 1-8. 2016.
[22] P.A.Arsaningtyas, M.A. Bijaksana. S.A. Faraby. Sistem Pencarian Ayat Al-Quran Berdasarkan KemiripanUcapan Menggunakan Algoritma Soundex dan Damerau-Levenshtein Distance. *Jurnal Linguistik Komputasional*. 1:2, pp. 58-64. 2018