Implementation of Winnowing Algorithm with Dictionary English-Indonesia Technique to Detect Plagiarism

Anton Yudhana, Sunardi
Department of Electrical Engineering
Universitas Ahmad Dahlan Yogyakarta,
Indonesia

Iif Alfiatul Mukaromah
Master of Informatics Engineering
Universitas Ahmad Dahlan Yogyakarta,
Indonesia

***Abstract*—The ease of obtaining information that is easy, fast, and cheap from all over the world through the internet network can encourage someone to take action plagiarism. Plagiarism is an intellectual crime that often occurs in the writing world where the perpetrators take the work of others without declaring the original source; if it continues to be left it will have a negative impact on the academic community and can be a chronic disease in the progress of a nation. At this time, the process of plagiarism detection is done manually and automatically using the help of technological developments (plagiarism detection), but the automatic checks available now mostly just check every letter character contained in the document, cannot check where the plagiarist takes a quote from a foreign language and changed in plagiarist language. Detection of plagiarism in this study will use a winnowing algorithm that has a function to check every character in two samples by hashing method that can generate fingerprint from two documents. While the dictionary method English-Indonesia change the writing from English to Indonesian language. This research will produce plagiarism detection using winnowing algorithm with English-Indonesian dictionary technique.**

Keywords—Plagiarism; winnowing algorithm; fingerprint; dictionary English-Indonesia

I. INTRODUCTION

Plagiarism is an intellectual crime and an unlawful act in which the offender attempts to take the work of another person either whole, in part or in small part without permission or without mentioning the original owner, so that the act of plagiarism is the same as the act of stealing [1], [2]. The act of plagiarism is not a new act but a practice that is often done (no stranger) again in the country of Indonesia and throughout the world that occurs in the academic world, the world of writing and in our society [1].

Plagiarism action occurs one of them is the ease of obtaining an information that can be accessed and taken any time example from the internet, the internet is the sophistication of technology that every year progressively rapid development [3], [4]. As the discovery of zalnur from the results of his research that there are two factors causing the occurrence of plagiarism among students that is the development of information technology is increasingly sophisticated and the burden of assignment given to the student is heavy enough so that many students choose an instant path by doing the act of plagiarism [1].

The act of plagiarism can not be allowed to develop especially in the world of education, this action can damage the academic community and can result in the decline of a nation because its critical mindset is not honed [2]. According to [2], a nation will experience decline because someone is lazy to think, no more development of science or new discoveries produced from the children of the nation. Even the plagiarism act indicates weak character education [1]. There needs to be action that can minimize the action of plagiarism in order not to become a habit of a nation. Even plagiarism is not only done by copy-pasting and changing every word for word with the same meaning, but some people do plagiarism by taking the work of foreign writing and converting it into another language (translated into plagiarist language) [3].

Prior research has implemented several algorithms that function as document fingerprints to detect plagiarism such as rabin-karp algorithm, winnowing algorithm, smith-waterman algorithm and so on [5]-[8]. However, these algorithms can only check every word in the document file and can not check the action of plagiarism done by taking other people's work in foreign language written by another language. In this case researchers will use the winnowing algorithm to detect plagiarism and this study can only check plagiarism in the category of translated plagiarism and plagiarism ideas in English sentences translated into Indonesian. Winnowing algorithm is an algorithm that has a function to check the similarity of words using hashing techniques and will be collated with dictionary English-Indonesia technique that serves to translate the writings from english to Indonesia.

Through this research, it is hoped that the application of plagiarism detection using winnowing algorithm with dictionary-english-indonesia technique can minimize the action of plagiarism especially in the category of translated plagiarism and idea plagiarism which is done by taking or quoting the writing in English which is translated into Indonesian.

II. RESEARCH METHOD

In this research will be focused on the implementation of winnowing algorithm with dictionary English-Indonesia technique to detect the existence of an action of plagiarism.

A. Categories of Plagiarism

The act of plagiarism is an act of stealing against the work of others because it does not reveal its original source [3]. Some of the things that plagiarists do is to quote or steal other people's work by paraphrasing quotations so that their actions are unknown.

According to B. Gipp and N. Meuschke (2011) in his research categorize the act of plagiarism based on the means used, among them [3]:

- Copy & paste plagiarism, copying entirely without any changes
- Disguised plagiarism, covering the copied parts, such as shake & paste, expansive plagiarism, contractive plagiarism, and mosaic plagiarism.
- Technical disguise, summarizing quotations to be subject to automatic detection by replacing letters with foreign letters.
- Undue paraphrasing, Paraphrasing quotations or alien thinking into the plagiarist language and hiding the original owner
- Translated plagiarism, translating quotations from one language to another.
- Idea plagiarism, using foreign ideas without declaring the source.

B. Winnowing Algorithm

Winnowing algorithm is an algorithm that has function as document fingerprint which is used to check the similarity of words in two documents by utilizing fingerprint concept with hashing technique [7], [9].

The winnowing algorithm is the exclusion of the rabin-karp algorithm by adding a window concept. Every word in the document will first be foxed in the hash form using the hash rolling formula by changing the characters in the document into ASCII code [10].

Here is the rolling hash formula that will be shown in (1) and (2).

$$H(C_1..C_l) = C_1.b^{(l-1)} + C_2.b^{(l-2)} + ... + C_{(l-1)}.b + C_l \quad (1)$$

$$H(c_2....c_{l+1}) = (H(c_1....c_l) - c_1 * b^{(l-1)}) * b + c^{(l-1)} \quad (2)$$

Where:

$H(C_1..C_l)$ = hash value
C_l = ASCII value of Character to -1 on string
l = string length
b = hash base value

The winnowing algorithm uses a certain window size and each window has a fingerprint that will be used to check the similarity of words on two documents or samples. The fingerprint to be selected is the smallest fingerprint, if there are two fingerprints in one window then select the rightmost fingerprint [9]-[11]. Fig. 1 is a concept of winnowing algorithm.

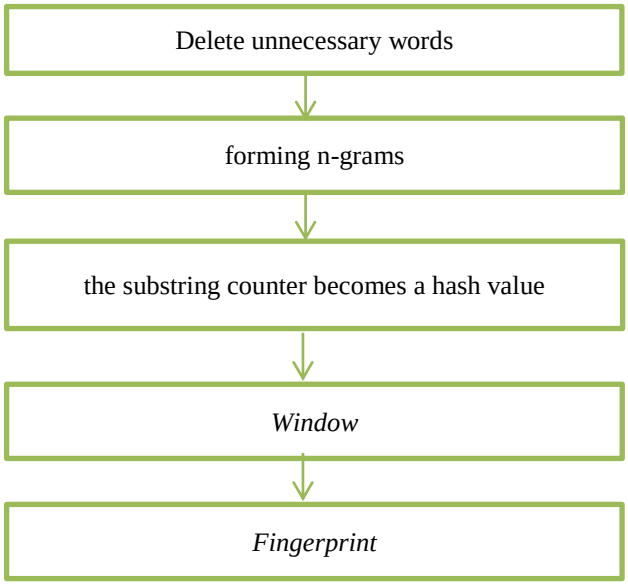

Fig. 1. The concept of Winnowing Algorithm.

The concept of the winnowing algorithm in Fig. 3 removes irrelevant characters, forming n-grams of length n, computing hash values, forming window values, and selecting hash values as document fingerfrint [6].

C. Dictionary English-Indonesia

Dictionary English-Indonesia is a dictionary to translate from English to Indonesian. This dictionary will be used to detect the action of plagiarism. The act of plagiarism in the writing world can not be separated from the use of foreign sentences that are translated into other languages by the actors of plagiarism to avoid the automatic detection tool.

This act of plagiarism includes the categories of Translated plagiarism and the idea of plagiarism. This category of plagiarism is very difficult to detect by means of plagiarism detection tool whose function is to check word equality on two documents by using fingerprint concept if not combined with Dictionary English-Indonesia. If in two documents there is an English vocabulary then the system will do the translation process into Indonesian before doing the process of checking the similarity in both documents. Dictionary English-Indonesia is highly dependent on databases containing English and Bahasa Indonesia vocabulary.

III. Proposed System

This section will discuss some of the supporting systems for making plagiarism detection using the dictionary English-Indonesia technique. The system will be designed as follows:

A. System Design Dictionary English-Indonesia

Dictionary English-Indonesia is an important thing to check an act of plagiarism that takes a foreign scientific work in English and changed into the Indonesia language.

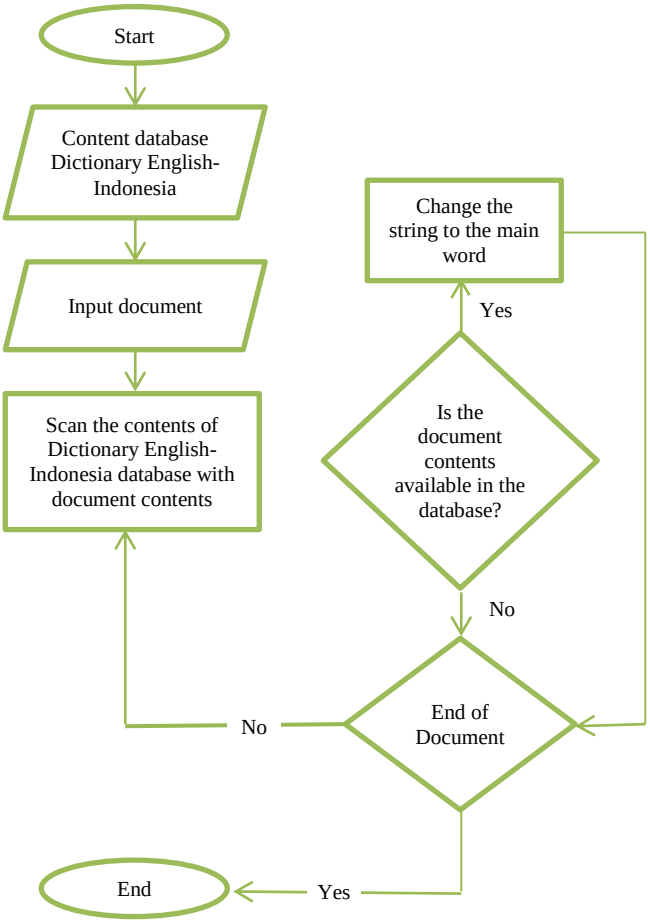

Fig. 2. Process dictionary English-Indonesia.

Based on the flowchart designated in Fig. 2 is a process of dictionary English-Indonesia, the outline is the process by which all the contents of the document will be scanned and will be matched with the existing word in the dictionary English-Indonesia database will then be modified based on the word available on the dictionary database English-Indonesia, if a string of text that match has an English word, the system will convert the text string into Indonesian language already available in the dictionary English-Indonesia database and the word will be included in the Winnowing process and will be reconciled. If a scanned and matched text string does not have an English word, then the process from dictionary English-Indonesia will not be performed. This stage will continue to repeat until the entire process of scan and string matching is complete.

Document I : saya makan apple

Document II : saya makan apel

With the dictionary English-Indonesia then the document has an English word will be changed into the Indonesia language based on dictionary English-Indonesia database.

Document I : saya makan apel

Document II : saya makan apel

After the process is complete then the sentence will go directly to the winnowing algorithm stage to checked again.

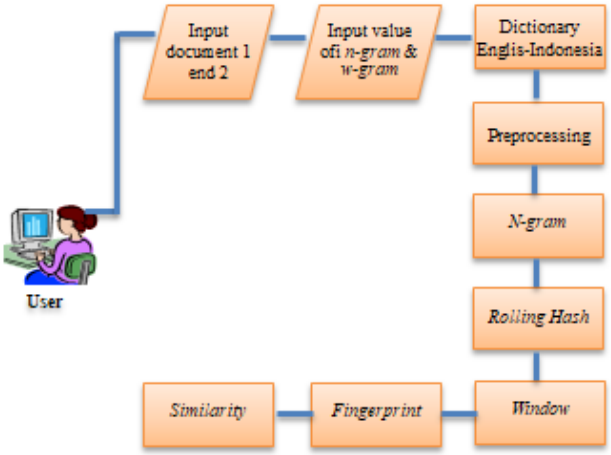

Fig. 3. Flow of Winnowing Algorithm with dictionary English-Indonesian technique for detection of plagiarism.

B. Designing Winnowing Algorithm System with Dictionary English-Indonesia Technique

In this research stages of winnowing algorithm with dictionary Englis-Indonesia technique. stages can be seen in Fig. 3.

Fig. 3 is a system flow for creating a tool for detecting plagiarism using a winnowing algorithm that serves to process two documents by adding the English-Bahasa Indonesia dictionary technique so that it can examine plagiarism in the category of translated plagiarism and plagiarism of ideas. At the beginning of the user interface must first enter the document to compare and include the document in comparison, then the user also need to specify and then enter the value of n-gram and w-gram (window) to be used as Fingerprint Search of both documents.

The process of generating fingerprints from two documents from the use of the winnowing algorithm with the English-Indonesian dictionary technique for detecting plagiarism shown in Fig. 3 is as follows:

1) Process dictionary English-Indonesia: The system will first perform the scanner process of the contents of two documents with the contents of the dictionary English-Indonesi database. If the user-input document has an English word available in the dictionary English-Indonesia database it will be scanned and will be converted into Indonesian, but if it does not have the English word available on the database it will proceed directly to the winnowing process.

2) Preprocessing: Removes irrelevant characters on a document [12].

a) Case Foling: The process of converting capital letters to lowercase in a document (a-z) [12].

b) Tokenizing: Removes unnecessary characters such as spaces [12].

3) N-gram: Serves to retrieve a token circuit or tangible character pieces along the length of n of a continuously inserted document (continuity) to shift according to the given offsite or end of a word or document [12], [13].

```
private function n_gram($word, $n) {
    $ngrams = array();
    $length = strlen($word);
    for($i = 0; $i < $length; $i++) {
            if($i > ($n - 2)) {
                    $ng = '';
                    for($j = $n-1; $j >= 0; $j--) {
                            $ng .= $word[$i-$j];
                    }
                    $ngrams[] = $ng;
            }
    }
    return $ngrams;
}
```

Fig. 4. N-gram program.

Fig. 4 is a PHP program from N-gram, the process will take a series of characters along the n-gram value specified by the user

4) Rolling hash is a hashing method that is used to find the hash values of the grams that have been formed and gives the ability to calculate values without repeating the entire string [14]. The hash value is a numerical value formed from the ASCII code [15]. Rolling hash formula can be seen in (1) and (2) above.

5) Window is the main process of winnowing algorithm that serves to categorize hash values that have been formed to produce fingerprint [14].

6) Fingerprint selects the smallest hash value of any given group in the window stage, if there are two or more smallest hash values then select the smallest right hash value.

7) Similarity: measures the similarity in two documents or samples [16]. Similarity to be used is Jaccard Coefficient is usually used to compare documents and calculate the similarity of two objects or documents [15], [17], [18]. Jaccard Coefficient can be seen in (3) [18].

$$Similarity\ (X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (3)$$

where:
X = Document 1
Y = Document 2

The size of a person performing a plagiarism act will be determined by similarity percentage level [5], [6]:

- 0% said the document has nothing in common.
- <15% say documents show plagiarism of small size / have little in common.
- 15-50% said the document is classified as moderate plagiarism.
- 50% mention the document including plagiarism with a large size.
- 100% stated that the document as a whole has something in common.

IV. Result

A. Implementation

Here is the implementation of winnowing algorithm with dictionary English-Indonesia technique to detect plagiarism.

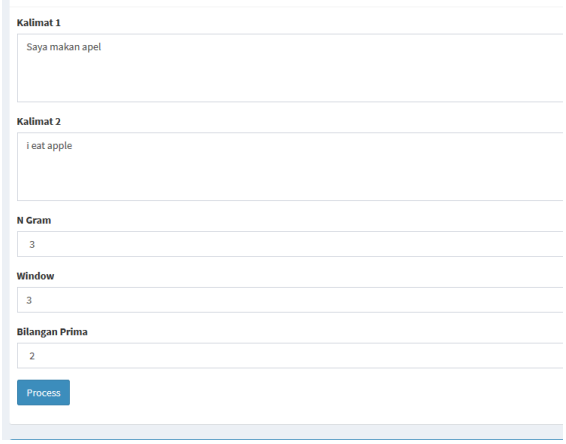

Fig. 5. Input Kalimat 1, Kalimat 2, n-gram and window.

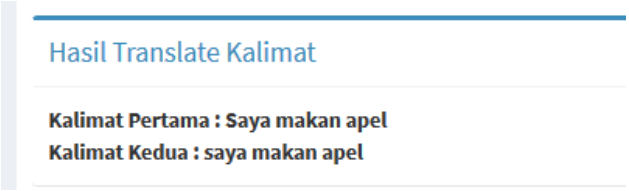

Fig. 6. Result dictionary English-Indonesia.

Fig. 5 is the first interface and the user must input the document you want to compare and must input the document that will be the comparison. The determination of n-gram and window values determines the value of similarity.

Fig. 6 shows the process of scanning on documents that have English, the document input process in the second sentence shown in Fig. 5 contains the sentence “i eat apple” and after going through the Dictionary English-Indonesia process “saya makan apel”.

Table I is an irrelevant character removal process of a document or text, whereby spaces are removed and convert capital letters into lowercase.

In Table II above can be seen that the value of n-grams 1 and 2 have similarities, so that hash values 1 and 2 also have the same overall value after doing dictionary English-Indonesia process.

TABLE I. Case Folding and Tokenizing

Kalimat 1	sayamakanapel
Kalimat 2	sayamakanapel

TABLE II. RESULT N-GRAM AND HASH

No	*N-gram 1*	*N-gram 2*	*Hash 1*	*Hash 2*
1	say	say	775	775
2	aya	aya	727	727
3	yam	yam	787	787
4	ama	ama	703	703
5	mak	mak	737	737
6	aka	aka	699	699
7	kan	kan	732	732
8	ana	ana	705	705
9	nap	nap	746	746
10	ape	ape	713	713
11	pel	pel	758	758

TABLE III. RESULT WINDOW

Window 1	**Window 2**
775 **727** 787	775 **727** 787
727 787 **703**	727 787 **703**
787 **703** 737	787 **703** 737
703 737 **699**	703 737 **699**
737 **699** 732	737 **699** 732
699 732 705	**699** 732 705
732 **705** 746	732 **705** 746
705 746 713	**705** 746 713
746 **713** 758	746 **713** 758

TABLE IV. FINGERPRINT

Fingerprint 1	727 703 703 699 699 699 705 705 713
Fingerprint 2	727 703 703 699 699 699 705 705 713

Table III is the result of grouping the hash value of a number of w-gram values, from the window process generating fingerprints on each document or sentence by selecting the smallest hash value.

Table IV shows the kalimat 1 and 2 have each fingerprint 9 and have the same value, from this process already visible, the data is a word that has the same meaning after the dictionary English-Indonesia process is done. to prove what percentage of similarity documents can be seen in Fig. 7.

The results from Fig. 7 show that both documents have a 100% similarity after performing the dictionary English-Indonesia process.

B. Trials

The test results of winnowing algorithm with dictionary English-Indonesia technique to detect plagiarism with plagiarism detection test using winnowing algorithm without dictionary English-Indonesia technique with n-gram and w-gram 3 values will be displayed in Table V.

Jumlah Fingerprints kalimat 1 = 9

Jumlah Fingerprints kalimat 2 = 9

Union (Gabungan) Fingerprints 1 dan 2 = 18

Intersection (fingerprints yang sama) = 9

(Union - Intersection) = 9

Prosentase Plagiarisme

Koefisien Jaccard = (Intersection / (Union-Intersection)) * 100

(9/9) * 100 = 100 %

Fig. 7. Similarity.

TABLE V. TEST RESULT OF PARAMETER

No	Input document	N-gram	W-gram	Similarity %	
				With DEI	Without DEI
1	Kesehatan itu penting Health is important	3	3	100%	0%
2	Perempuan itu sangat cantik dan cerdas Dokter itu very beutiful and smart	3	3	66,67%	8%
3	Plagirsm merupakan salah satu problem dalam dunia academic Plagiarisme merupakan salah satu permasalahan dalam dunia akademik dan permasalahan bagi bangsa	3	3	49,41%	33,68%
4	Algoritma winnowing berfungsi sebagai document fingerprint dengan teknik hashing Algoritma winnowing berfungsi sebagai dokumen sidik jari dengan teknik hashing	3	3	100%	82,19%
5	Dua dokumen tersebut memiliki nilai kesamaan yang sama Dua dokumen tersebut memiliki nilai similarity yang berbeda	3	3	93,48%	59,65%

The above Table V is the result of the parametric trials of the winnowing algorithm with the dictionary English-Indonesia technique and the winnowing algorithm without dictionary English-Indonesia. Can be seen from Table V the influence of winnowing algorithm with dictionary English-Indonesia technique can give high similarity value (high accuracy) than not using Dictionary English-Indonesia technique. In Table V the number 1 data entered in sentences 1 and 2 are different data but have the same meaning. In sentence 1 data entered with the Indonesian language, while in sentence 2 data entered with English. With dictionary English-Indonesia technique then the system will change every word in English into Bahasa Indonesia, after that new system will do re-checking by using winnowing algorithm so that in Table V number 1 yields 100% similarity value. Whereas if detected using only winnowing algorithm alone without dictionary English-Indonesia technique, the system will only check every document in inputkan by user without any change, so that in Table V number 1 yields the value of similarity 0% because the document entered user in sentence 1 and 2 has a very different character though it has the same meaning.

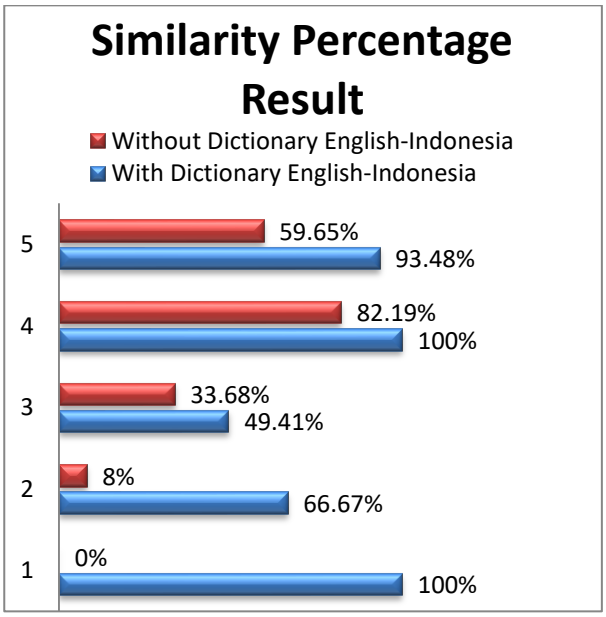

Fig. 8. Similarity percentage result.

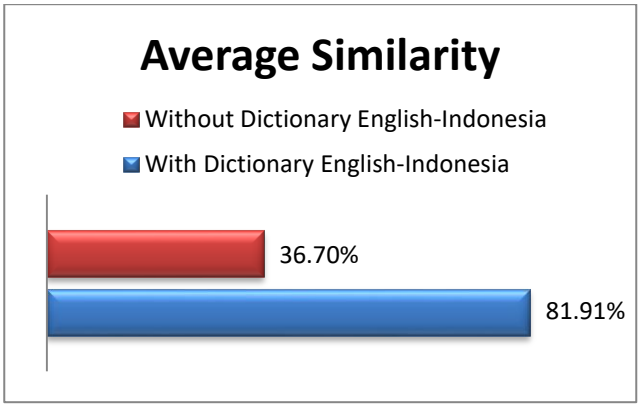

Fig. 9. Average similarity.

Fig. 8 shows the similarity percentage results of the experiments performed in Table V. It can be seen from the Fig. 8 similarity values generated using the dictionary English-Indonesia technique has a high accuracy compared to similarty value without using dictionary English-Indonesia. This is because the algorithm that serves as a document fingerprint can only check every character contained in the contents of the document entered, while the action of plagiarism is not just copy paste, but many categories in the act of plagiarism. one of which is to avoid the plagiarism detection tool usually the perpetrator hides the copied part, summarizes the copied part, paraphrases the part copied by plagiarist style and translates the part copied from the foreign work into the plagiarist language.

In Fig. 9, it can be seen that the use of dictionary English-Indonesian technique has a very big influence on the accuracy of the position of the sentence. Can be seen from Fig. 9, there is a difference of accuracy obtained up to 45.21%. This proves that winnowing algorithm with dictionary English-Indonesia technique is very useful in the accuracy of sentence position.

V. Conclusion

From the above research can be understood that the algorithm winnowing with dictionary English-Indonesian technique in plagiarism detection tool is very important to prevent the action of plagiarism in the category of translated plagiarism and idea plagiarism. Plagiarism detection algorithms such as winnowing, rabin-karp algorithm and so on only have a function for pattern matching according to documents or samples that have been inputkn users. The system is unable to check sentences, quotations, or paragraphs that have been paraphrased, hidden, and transransced by actors of plagiarism. Winnowing algorithm with dictionary English-Indonesia technique to detect plagiarism very well is used to minimize the action of plagiarism in the category of translated plagiarism and idea plagiarism, the dictionary English-Indonesia technique also increases the value of similarity between documents.

It is expected that the results of this study can be continued as a follow-up study by researchers themselves and by other researchers. For example, plagiarism detection tool using winnowing algorithm with dictionary Indonesia-English, Arab-English, Mandarin-English and others, to detect plagiarism in category of translated plagiarism and idea plagiarism. In addition, the development of a plagiarism detection tool using the winnowing algorithm can be further developed using the Rabin-karp algorithm, the smith-waterman algorithm, and/or the combination of some of these algorithms.

References

[1] M. Zalnur, "Plagiarisme Di Kalangan Mahasiswa Dalam Membuat Tugas-Tugas Perkuliahan Pada Fakultas Tarbiyah Iain Imam Bonjol Padang," AL-Ta lim, vol. 19, p. 55, 2012.

[2] A. Wibowo, "Mencegah dan Menanggulangi Plagiarisme di Dunia Pendidikan," Kesmas J. Kesehat. Masy. Nas., vol. 6, no. 5, pp. 195–200, 2012.

[3] Sunardi, A. Yudhana, and I. A. Mukaromah, "Perancangan aplikasi deteksi plagiarisme karya ilmiah menggunakan algoritma winnowing," in Prosiding SNSebatik, 2017, vol. 1, no. 1, pp. 27–32.

[4] N. F. Ulfa, M. Mustikasari, and I. Bastian, "Pendeteksian tingkat similaritas dokumen berbasis web menggunakan algoritma winnowing,"

in Konferensi Nasional Teknologi Informasi dan Komunikasi (KNASTIK), 2016, pp. 194–203.

[5] A. Yudhana and A. D. Djayali, "Implementation of Pattern Matching Algorithm for Portable Document Format," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 11, pp. 509–512, 2017.

[6] A. Yudhana, A. D. Djayali, and Sunardi, "Sistem Deteksi Plagiarisme Dokumen Karya Ilmiah dengan Algoritma Pencocokan Pola," JURTI, vol. 1, no. 2, pp. 178–187, 2017.

[7] R. K. Wibowo and K. Hastuti, "Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Teks pada Tugas Akhir Manusia," Techno.COM, vol. 15, no. 4, pp. 303–311, 2016.

[8] R. V. Imbar et al., "Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks," J. Inform., vol. 10, no. 1, pp. 31–42, 2014.

[9] A. T. Wibowo, K. W. Sudarmadi, and A. M. Barmawi, "Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents," 2013 Int. Conf. Inf. Commun. Technol. ICoICT 2013, no. March, pp. 128–133, 2013.

[10] G. Wu, M. Zhao, L. Han, and S. Li, "A Fingerprint Feature Extraction Algorithm based on optimal Decision for Text Copy Detection," Int. J. Secur. Its Appl., vol. 10, no. 11, pp. 67–78, 2016.

[11] K. T. Tung, N. D. Hung, L. Thi, and M. Hanh, "A Comparison of Algorithms used to measure the Similarity between two documents," Int. J. Adv. Res. Comput. Eng. Technol., vol. 4, no. 4, pp. 1117–1121, 2015.

[12] E. Nugroho, "Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp Skripsi," in Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Brawijaya Malang, 2011.

[13] E. A. Lisangan, "Implementasi n-gram Technique dalam Deteksi Plagiarisme pada Tugas Mahasiswa," J. Temat., vol. 1, no. 2, pp. 24–30, 2013.

[14] M. Ridho, "Rancang Bangun Aplikasi Pendeteksi Penjiplakan Dokumen Menggunakan Algoritma Biword Winnowing," in Teknik Informatika Universitas Islam Negeri SLTAN Syarif Kasim Pekan Baru Riau, 2013.

[15] K. Rinartha, "Simple Query Suggestion untuk Pencarian Artikel menggunakan Jaccard Similarity," J. Ilm. Rekayasa dan Manaj. Sist. Inf., vol. 3, no. 1, pp. 30–34, 2017.

[16] S. A. Djayali, A. Yudhana, "Pendeteksian Plagiarisme dengan Sistem Pengukuran Similartas pada Dokumen Karya Ilmiah Menggunakan String Matching Rabin-Karp," in Cyber Learning & It Computer Karawang, 2016, vol. 1, no. 1.

[17] M. Fadelillah, I. Much Ibnu Subroto, and D. Kurniadi, "Sistem Rekomendasi Hasil Pencarian Artikel Menggunakan Metode Jaccard ' s Coefficient," J. Transistor Elektro dan Inform. (TRANSISTOR EI), vol. 2, no. 1, pp. 1–14.

[18] S. Sugiyamto, B. Surarso, A. Sugiharto, and S. A, "Analisa Performa Metode Cosine dan Jacard pada Pengujian Kesamaan Dokumen," J. Masy. Inform., vol. 5, no. 10, pp. 1–8, 2016.