 ISSN: 1978-0524

Indonesia words detection using fingerprint winnowing algorithm

Sunardi[a,1],, Anton Yudhana[b,c,2], Iif Alfiatul Mukaromah[b,3*]

[a ,b]Department of Electrical Engineering. Universitas Ahmad Dahlan, Indonesia
[c]Master of Informatics Engineering, Universitas Ahmad Dahlan, Indonesia
[1]Sunardi@mti.uad.ac.id, [2]beyudhana@ee.uad.ac.id, [3]ciifam1604@gmail.com[*]
** corresponding author*

ABSTRACT

The action of plagiarism is bad deeds that violate the code of ethics in the world of writing that often occur in the academic environment, the lack of interest in reading and creativity of the nation is one of the factors that cause obstruction of the progress of a nation. Besides the plagiarism action bad for good name at a college which is caused by the plagiarist. This needs to be minimized or prevented so that the problem does not happen again. the use of fingerprint in winnowing algorithm can solve the problem. winnowing algorithm is an algorithm used to match a word for word with hashing method, the use of n-gram and w-gram is very influential on result of similarity.

Keywords:
Plagiarism
Fingerprint
Algoritma Winnowing
n-gram
w-gram

I. Introduction

Internet is the evidence of advanced technological developments, there are lots of positive side of internet, for example, internet has made human easier and faster to get an information, and also facilitate humans in completing tasks such as scientific work. Beside the positive side Internet also has a negative sides if used by people with bad intentions, for example, many users used internet for doing article plagiarism, especially on scientific work related articles. This might caused by the society limited reading interest and time so that there is encouragement to perform the act of plagiarism[1].

Plagiarism is a unethical behaviour in the writing world. It acknowledges or hijacks the results of other people's thoughts, then acknowledges that the work or thought result is the result of its thinking. The act of plagiarism can lead people to become more ignorant because they are not trying to generate new thinking, because there is no desire to develop or generate new ideas for better development. The act of plagiarism is also a form of violation of government regulations [2]. According to [2] these actions will bring down the morale of the nation due to the un-creative community. The act of taking the work of another person is allowed as a reference in the work of scientific work or in doing a research with the condition to include the source, and better if we develop research or the work of others become much better and can follow the development of the era. The action of plagiarism should be prevented and must be seriously addressed, especially in the academic filed so as not to affect the integrity of academic community as well as the progress of a nation [2].

Based on the above problems Winnowing algorithm can be a solution to prevent or minimize the action of plagiarism, winnowing algorithm is the process of checking the fingerprinting to detect the existence of a plagiarism by using hashing technique [3]. The winnowing algorithm is sufficient to prevent or minimize the action of plagiarism by forming N-grams and w-grams with small values because the algorithm will check the strings along the specified n-grams and w-grams.

II. Literature Review

A. Plagiarism

Plagiarism is taking other's work (opinion and so on) and acknowledging the work itself, for example publishing papers or other people's thoughts on its own behalf, without mentioning its

DOI: 10.26555/jifo.v13i1.a8452

source; plagiarism ". The person performing the plagiarism act is called a plagiarist, ie "one who takes a work (opinion and so on) of others and is published as his own; copycat [4].

Plagiarism is the hijacking of the works of others and acknowledges them as his own works [5]. According to [5] results or the work of others who deliberately imitated by the plagiarist and recognized as his work. Kramer Et Al (1995) and Wray (2006) say that the occurrence of the act of plagiarism is when a writer quotes pieces of work such as ideas, opinions, conclusions, sentences, findings, data and words of others so that the reader considers the work is the result of author's thought [6]. According to [6] who quotes or takes the work of others as a reference source of research or work in progress, it is permissible, with the condition to include the source.

According to B. Gipp and N. Meuschke (2011), that the act of plagiarism is not merely a copy-paste but many plagiarists are hijacking the work of others by changing the words of each word that has the same meaning, hijacking the work of others from a foreign language and converted into his own language, hijacking foreign ideas without including its source [7].

B. Winnowing Algorithm

Winnowing algorithm is an algorithm that works to check the equality of a word or document (document fingerprint) to detect the existence of a plagiarism, by detecting on the smallest parts of a large number of document texts, which will then be processed to produce the result of a collection of hashes and these hash values is used as a fingerprint in comparing between documents [3] [8].

The winnowing algorithm is technic ally an extension of rabin-karp fingerprint Algorithm implementation by adding the window method [9]. According to [9] states winnowing algorithm in the process is almost the same as rabin-karp algorithm, only in winnowing algorithm added window method as the process of winnowing algorithm.

Stages in the winnowing Algorithm:

- Preprocessing: removes irrelevant words in text documents [10]. Preproccesing methods include:
  1. Case folding: all text input will be changed to lowercase (a to z) [10].
  2. Tokenizing: the process of separating words based on each word that compiles them [9].
  3. Filtering: the result of tokenizing phases will be continued by taking important words, this step is done by removing stoplist / stopword (words that are not descriptive, for example the word "from, which, and, or" and so on [11]. and can reduce processing time [12].
  4. Stemming: This stage will do the separation of words into the word or the original form, this stage is done to find the root word of each word on filtering [10].
- N-gram is a series of tokens of length n. N-gram is a method that serves to take pieces of letters and numbers as much as n [9].
- Rolling hashes are a hashing method that is used to find the hash values of the grams that have formed and give the ability to calculate values without repeating the entire string. The hash value is a numerical value that is formed from the ASCII code (American Standard Code for Information Interchange) [9][13].

  Formula (1) and (2) explains how rolling hash formula works:

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

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

  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
  Initial calculations on the earliest n-gram sequences are calculated using the formula number 1, and the next n-gram circuit until the last gram sequence is calculated using the formula number 2, so the process will be much faster because it does not count anymore from the beginning.

- Window is the core process of winnowing, the window is used to form the resulting substrings in the hash rolling process along w-gram [9].
- Fingerprint works for matching the text of the text or plagiarism [1].
- Similarity is the percentage of similarity levels between documents. In this section there will be a matching process between two documents from the smallest fingerprint selection process that has passed the winnowing process [1].

$$(1, 2)=\frac{1\cap 2}{1\cup 2} \text{ X } 100\% \quad (3)$$

The technique built in determining the similarity value (documented) consists of 3 part [14]. Among others:

1. Distance-based similarity measure is to measure the level of similarity of two objects in terms of geometric distance from the variables enclosed within the two objects. Distance-based Similarity methods include Minkowski Distance, Manhattan / City Block Distance, Euclidean Distance, Jaccard Distance, Dice's Coefficient, Cosine Similarity, Levenshtein Distance, Hamming Distance and Soundex Distance.
2. Feature-based similarity measure is the calculation of similarity level by representing the object into the form of features that want to be compared. Feature-based similarity is widely used in classifying or pattern matching for images and text.
3. Probabilistic-based similarity measure, which calculates the level of similarity of two objects by representing two sets of objects that are compared in the form of probability. Includes Leibler Distance Kullback and Posterior Probability.

Size of similarity [15]:
1. 0% : the results of two text documents have nothing in common
2. <15% : results from two text documents have little in common
3. 15-50% : included in the category of moderate plagiarism
4. 50% : the results of both text documents detect the presence of plagiarism
5. 100% : the text document is plagiarism

III. Method

A. Research Method

This research used waterfall method, which means the stage will be done after the completion of the previous stage.

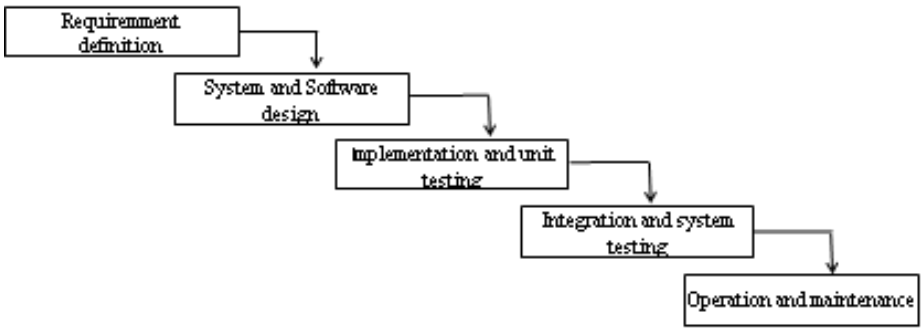

Fig. 1. Waterfall Method

Figure 1 is the waterfall method step according to Sommerville (2011: 30-31). Here are the steps of waterfall [7]:

- Requirements definition: Seeking needs or referral results to assist in the development of research to be researched, such as finding or collecting reference sources for developing a plagiarism detection system.
- System and Software Design: This process transforms the above needs into a representation into a "blueprint" software before coding begins. At this stage is analyzing the system of plagiarism detection
- Implementation and Unit Testing: Implementation process of the program has been designed.
- Integration and Sytem Testing: This stage is the application testing process performed. So at this stage apply winnowing algorithm for the operation of applications that have been designed.
- Operation and Maintenance: maintenance of software or applications that have been made, and is expected there is a development of the application for the system designed to be better.

B. System designd

The design of the system used to minimize the act of plagiarism in the world of writing by checking every word for word, each string in a sentence or document with a hash calculation, will be illustrated in figure 2.

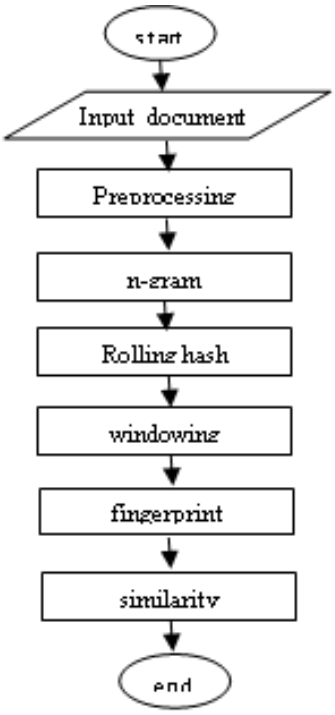

Fig. 2. Process winnowing

Figure 2 illustrates that the winnowing process requires the input of two text documents to test the similarity of two text documents by going through several stages of the process so that it can produce what the similarity is. Preprocessing is a step that will be done immediately after two documents tesks to be matched already in inputkan that this stage will throw away word or unneeded characters. Furthermore, the n-gram stage of one will affect the final process, the lower the similarty result means n-gram in the inputkan large, the higher the similarity results then the n- gram included is small. In the next step every word or sentence that has been in form along the gram will be converted into ASCII code, so that will be done rolling hash process. The next stage of windowing or w-gram is to form hash values along w-gram, w-gram formation is very influential on similarity results, the same principle with the formation of n-gram. The smallest hash value of the windowing will be used as the

fingerprint of both documents. In the latter stages the two text documents will be known to match the degree of similarity of the selected fingerprint.

IV. Results and Discussion

Implementation of Indonesian words detection using fingerprint Winnowing algorithm with different n-gram and w-gram formation.

A. Input Document

at the beginning of the interface, the user must input text or .txt-shaped documents into sentences one and two to be tested for similarity and then processed.

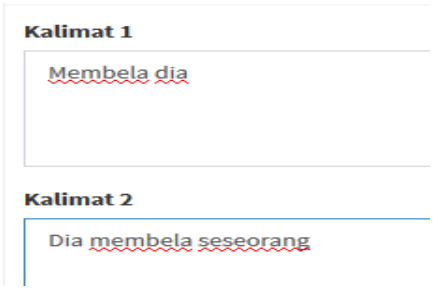

Fig 3. Input Document

Figure 3 shows a document input process. In kalimat 1 containing "Membela dia" and kalimat 2 contains "Dia membela seseorang", then it will be processed first at the preprocessing stage before on the formation of n-gram.

B. Preprocessing

This section removes unnecessary or eliminates irrelevant characters, such as converting uppercase to lowercase, removing spaces, marks (dots, commas and so on).

Table 1. Case Folding and Tokenizing

Kalimat 1	membeladia
Kalimat 2	diamembelaseseorang

Table 1 shows the result of the input of a word that has been through the preprocessing stage, in kalimat 1 after going through the preprocessing stage, the sentence becomes "membeladia", the whitespace is lost, the capital letter is changed to lower case, so also in kalimat 2. so the process will directly go to the n-gram formation stage.

C. The Formation of N-gram

This process will cut the number of letters n, meaning the value of n we can determine its own.

Table 2. Formation of N-gram with value 3 and 5

N-gram (value)	**N Gram 1**	**N Gram 2**
	Membela dia	**Dia Membela Seseorang**
3	mem emb mbe bel ela lad adi dia	dia iam ame mem emb mbe bel ela las ase ses ese seo ora ran ang
5	Membe embel mbela belad eladi ladia	Diame iamem amemb membe embel mbela belas elase lases asese seseo eseor seora eoran orang

Table 2 shows results from the formation of n-grams with values of 3 and 5. In n-gram with a value of 3 in N Gram 1 after being formed into " mem emb mbe bel ela lad adi dia ", and on N Gram 2 after formed with value 3 being " dia iam ame mem emb mbe bel ela las ase ses ese seo ora ran ang ". N-gram with value 5 on N Gram 1 becomes " Membe embel mbela belad eladi ladia", and in N Gram 2 with a value of 5 being "Diame iamem amemb membe embel mbela belas elase lases asese seseo eseor seora eoran orang". the formation of n-gram with value 3 will check each string to be shorter and more often the string will be checked, while the n-gram with the value of 5 strings will be checked for longer and the strings will be more rarely checked. so it will affect the similarity results. Proper use of n-grams will result in a high similarity value, the smaller the value of n-gram, the higher the similarity and the greater the value of n-gram, the lower the similary result.

D. Rolling Hash

This technique is used to get hash values from a series of grams that have been formed and hash rolling able to calculate values without having to count / repeat from scratch [8].

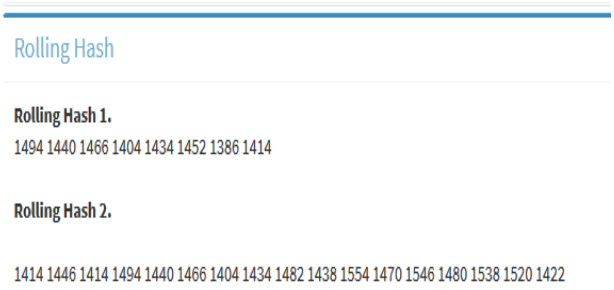

Fig 4. Process Rolling Hash

Figure 4 shows a calculation result using a rolling hash formula by first converting the letter or character that has been n-formed into the ASCII code is the American standard code for information exchange. The explanations in sentences 1 and 2 are listed in Table 3.

Table 3. Hash Value of each Substring

N-gram 3						**N-gram 5**					
Kalimat 1		**Kalimat 2**				**Kalimat 1**		**Kalimat 2**			
Substring	***Hash***	***Substring***	***Hash***	***Substring***	***Hash***	***substring***	***Hash***	***substring***	***Hash***	***substring***	***Hash***
mem	1494	dia	1414	ase	1438	Membe	6570	Diame	6294	asese	6414
emb	1440	iam	1446	ses	1554	embel	6380	iamem	6406	seseo	6842
mbe	1466	ame	1414	ese	1470	mbela	6490	amemb	6288	eseor	6552
bel	1404	mem	1494	seo	1546	belad	6204	membe	6570	seora	6834
ela	1434	emb	1440	eor	1480	eladi	6346	embel	6380	eoran	6528
Lad	1452	mbe	1466	ora	1538	ladia	6422	mbela	6490	orang	6798
Adi	1386	bel	1404	ran	1520			belas	6234		
dia	1414	ela	1434	ang	1422			elase	6398		
		las	1482					lases	6562		

Table 3 describes the substrings of "mem", after being converted into ASCII code and calculated using a rolling hash formula, so the hash is 1494, and an "emb" value with hash value 1440 and so on.

E. Windowing

After the hash value is generated then it will be continued in the windowing process by grouping the hash that has been formed in the rolling hash process. The smaller the size the higher the similarity and the greater the wgram the lower its similarity. This process is similar to the process of the n-gram process, a very biased way to similiraty results

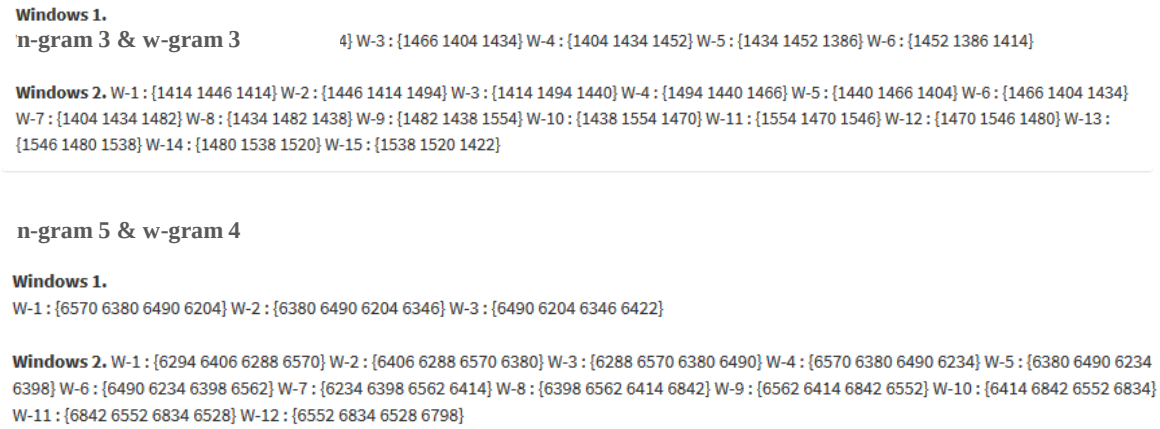

Fig 5. Establishment of w-gram value with value 3 and 4

Figure 5 shows the result of the formation of w-gram with values 3 and 4. The hash formed along w-gram is used for fingerprint retrieval as long as the user specifies, and takes the smallest hash value of the w-gram that has been formed. The formation of n-gram 3 and w-gram 3 in the formation of the first window is 1494 1440 1466, then the smallest hash value is selected 1440 and so on.

F. Fingerprint

This stage selects the smallest fingerprint of the hash that has been formed and takes the smallest fingerprint from the far right if the fingerprint is more than one.

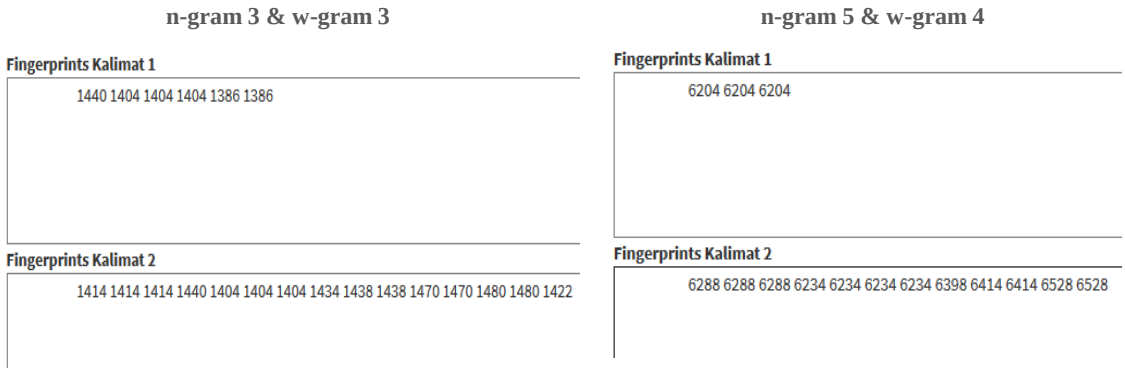

Fig 6. Process Fingerprint

Figure 6 shows the fingerprint process of both documents or sentences with different n-gram and w-gram formation. The formation of n-gram and w-gram with value 3 yields 6 fingerprints on fingerprint kalimat 1 and 15 fingerprint on fingerprint kalimat 2. formation of n-gram 5 and w-gram 4 produces 3 fingerprint on fingerprint kalimat 1 and 12 fingerprint on fingerprint kalimat 2 The resulting fingerprint will be the determinant of equality in two documents or sentences.

G. Similarity

the last step is the matching process of two documents that have been selected the smallest fingerprint on each document.

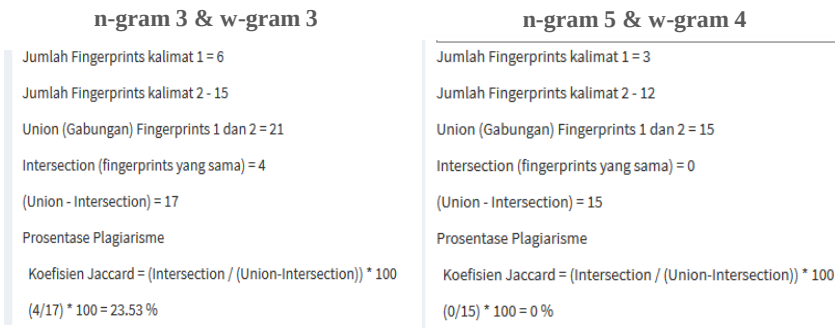

Figure 7 hows the result of similarity in two documents. The two documents having different

levels of similarity. In the formation of n-gram 3 and w-gram 3 both documents have a 23.53% census. while in the formation of n-gram 5 and w-gram 4 both documents have no similarity that is 0%. So to find out the similarity level of two documents is to divide the same fingerprint with different fingerprint, then multiplied by 100, so that the equality level of the two documents will be known. The formation of n-grams and w-grams greatly affects the similarity results.

Acknowledgment

The author would like to say a lot of praise to Allah SWT who has provided ease for authors in completing this journal for the smooth task of seminar courses and Scientific Writing. Also thanks to friends who have supported the author in completing this research.

References

[1] I. Mukaromah. Sunardi, A. Yudhana, "Perancangan Aplikasi Deteksi Plagiarisme Karya Ilmiah Menggunakan Algoritma Winnowing," in *Seminar Nasional Serba Informatika*, 2017.

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

[3] 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.

[4] Kamus Besar Bahasa Indonesia, "Arti kata plagiat - Kamus Besar Bahasa Indonesia (KBBI) Online." [Online]. Available: https://kbbi.web.id/plagiat. [Accessed: 26-Dec-2017].

[5] Salmuasih and A. Sunyoto, "Implementasi Algoritma Rabin Karp untuk Pendeteksian Plagiat Dokumen Teks Menggunakan Konsep Similarity," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2013, pp. 23–28.

[6] 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.

[7] M. Rosmiati, "Analisis dan Perancangan E-Service untuk Pelanggan pada Jaya Bersama Konveksi," *IJSE- Indones. J. Softw. Eng.*, vol. 1, no. 1, 2015.

[8] 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.

[9] 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.

[10] R. Y. Dillak, F. Laumal, and L. J. Kadja, "Sistem Deteksi Dini Plagiarisme Tugas Akhir Mahasiswa menggunakan Algoritma N-Grams dan Winnowing," *J. Ilm. FLASH*, vol. 2, no. 1 juni, 2013.

[11] 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.

[12] A. Kurniawan, F. Solihin, and F. Hastarita, "Perancangan dan Pembuatan Aplikasi Pencarian Informasi Beasiswa dengan Menggunakan Cosine Similarity," *J. SimanteC*, vol. 4, no. 2 Desember, 2014.

[13] Jody, A. T. Wibowo, and A. Arifianto, "Analisis dan Implementasi Algoritma Winnowing dengan Synonym Recognition pada Deteksi Plagiarisme untuk Dokumen Teks Berbahasa Indonesia," *e-Proceeding Eng.*, vol. 2, no. 3, pp. 7674–7683, 2015.

[14] 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.

[15] 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.