Science of Information & Technology Applied (SINTA)
Vol. 1 No. 1, March 2025: 16-32
E-ISSN : XXXX-XXXX, P-ISSN: XXXX-XXXX
https://ejournal.bacadulu.net/index.php/sinta

Classification of Document Similarity Using Winnowing Algorithm with Jaccard Coefficient Approach

Uzdha Zachrias[1], Wawan Gunawan[2*]
[1,2] Universitas Mercu Buana, Jakarta, Indonesia

(*) Corresponden Author: wawan.gunawan@mercubuana.ac.id

Article Info:

Abstract

Keywords:
Jaccard Coefficient;
Plagiarisme;
Similarity;
Stopword;
Winnowing;

Article History:
Received : dd-mm-yyyy
Revised : dd-mm-yyyy
Accepted : dd-mm-yyyy

Article DOI :
http://dx.doi.org/

In the era of information technology advancement, easy access to various sources of information through the internet has changed the way students conduct research. While it provides significant benefits, this convenience also brings the problem of plagiarism, which is a detrimental act in the academic world. Plagiarism is the act of copying or taking ideas from someone else's work without giving proper credit, which is contrary to academic guidelines. This research aims to develop an effective plagiarism detection system that is in accordance with the Indonesian language. This system uses a Winnowing algorithm with a Jaccard Coefficient approach and a technique of eliminating non-descriptive words (stopwords) in Indonesian. Samples of documents in Indonesian were taken from the final project of Mercu Buana University students. The data is collected from the university's repository and will be analyzed to measure the level of similarity between documents and the performance of the Winnowing algorithm in detecting plagiarism. The results of this study show that the development of a plagiarism detection system using the Winnowing algorithm and the Jaccard Coefficient approach with an n-gram value of 7 succeeded in achieving optimal results with precision, recall, and accuracy results reaching 100%. The similarity index detection system is able to provide accurate and relevant results on Indonesian documents.

How to cite :

INTRODUCTION

Technological advances have had a significant impact on the field of science. One of them is the ease of access to various information stored in digital documents via the internet [1]. This convenience greatly helps students in their research efforts by facilitating quick access to reference materials. Nonetheless, this also poses a problem because this ease of access can lead to unethical behavior, such as improper citations via copy-paste, which ultimately results in acts of plagiarism, which are contrary to established academic guidelines.

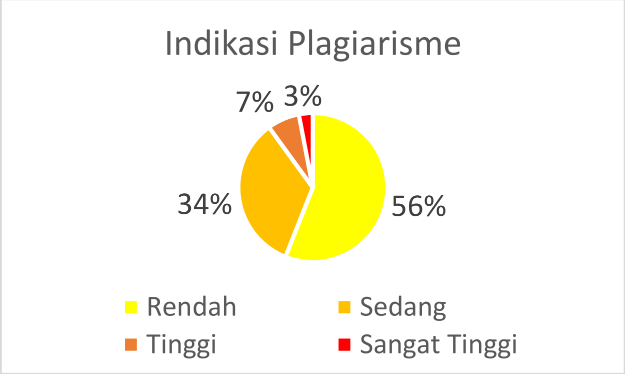
Plagiarism is a serious problem in the academic and research world. In accordance with the regulations of the Minister of National Education No. 17 of 2010 concerning the

e-ISSN: xxxx-xxxx
p-ISSN: xxxx-xxxx

Prevention and Prevention of Plagiarism in Higher Education, it is stated that plagiarism is an activity that is intentional or unintentional to assess a scientific work by quoting part or all of the scientific work or scientific work of another party. Plagiarism has created a bad climate, especially for the world of education. This action can kill ideas and ideas and lower a person's level of creativity [2].

Plagiarism is a serious problem in the academic and research world. In accordance with the regulations of the Minister of National Education No. 17 of 2010 concerning the Prevention and Prevention of Plagiarism in Higher Education, it is stated that plagiarism is an activity that is intentional or unintentional to assess a scientific work by quoting part or all of the scientific work or scientific work of another party. Plagiarism has created a bad climate, especially for the world of education. This action can kill ideas and ideas and lower a person's level of creativity [3]. The diagram can be seen in figure 1.

Figure 1. Indications of Plagiarism at the Student Level



To overcome this problem, efforts to check and control plagiarism are carried out by the university, by requiring students to attach the results of the similitary index test (an index of similarity of written writings that are declared plagiarized), in every preparation of thesis, thesis and dissertation. The scientific work can be declared as plagiarism if the percentage of the similitary index is high. However, in writing the scientific paper, it is indeed required to use and cite the opinions of experts and scientific literature that is relevant to their research. Based on the proportion and similarity level of the document, plagiarism is classified as follows: Light plagiarism, if the similarity rate is below 30%. Plagiarism is moderate, if the similarity rate is between 30% and 70%. Plagiarism is severe, if the similarity level is above 70% [4].

Detecting plagiarism can be done manually, which involves human intervention. However, this method has the disadvantage of being time-consuming and labor-intensive and prone to inconsistencies due to human emotional factors. Therefore, academics are actively working to develop systems that are capable of detecting plagiarism with a high degree of accuracy [2].

Actually, there are many software and websites that can be used to check the similitary index in document text, but it is not suitable for papers or scientific works written in Indonesian, because it is designed for English text. Therefore, it is necessary to design a similitary index detection application system that is more in accordance with text documents whose writing structure is in Indonesian [4].

Previous research conducted by Sugiono compared two algorithms to measure the level of effectiveness in creating a similarity detection system. There are two algorithms tested in the study to detect plagiarism by string matching, namely the Rabin-Karp

Algorithm and the Winnowing Algorithm. As a result, the Winnowing algorithm excels in terms of accuracy and processing time. The results of the experiment showed that the Winnowing algorithm was more effective in detecting similarities in text in document form [5].

Another study conducted by Sunardi entitled "Implementation of Plagiarism Detection Using the N-Gram and Jaccard Similarity Methods to Winnowing Algorithms" produced a similarity rate of up to 100% in plagiarism detection. This result was obtained using the Jaccard Similarity method with an n-gram value of 3. Compared to the k-gram method which produced 83% similarity, it shows that the Jaccard Similarity method has strong potential in plagiarism detection [6]. This indicates that this method is reliable for detecting plagiarism in the document or sample being studied.

Furthermore, another study was conducted using the Winnowing algorithm to identify plagiarism in text-based documents discussing Indonesia. The research conducted by Nurdiansyah involves two main stages. The first stage is the creation of document fingerprinting using the Winnowing algorithm, and the second stage uses Jaccard coeffcient to calculate the degree of similarity between documents. The results of this study show that the Winnowing algorithm produces high accuracy in detecting plagiarism. However, the study also evaluated the addition of elimination of non-descriptive words such as "which", "and", "in", "from" [7].

Based on the description above, the researcher is interested in developing a similitary index detection system, which applies the Winnowing algorithm with the Jaccard Coefficient approach and the technique of eliminating non-descriptive words (stopwords) in Indonesian. This system is expected to provide a more accurate and relevant solution in detecting plagiarism in Indonesian documents. This study is entitled "Classification of Document Similarity Using Winnowing Algorithm with Jaccard Coefficient Approach".

METHODS

Winnowing

Based on the description above, the researcher is interested in developing a similitary index detection system, which applies the Winnowing algorithm with the Jaccard Coefficient approach and the technique of eliminating non-descriptive words (stopwords) in Indonesian. This system is expected to provide a more accurate and relevant solution in detecting plagiarism in Indonesian documents. This study is entitled "Classification of Document Similarity Using Winnowing Algorithm with Jaccard Coefficient Approach".

Rolling Hash

Rolling hash rolling is a method used to generate a hash value of n-grams. The resulting hash is a numerical representation derived from the ASCII code. This hash value will be used to compare n-grams among different documents [8].

Formula rolling hash:

$$\mathrm{H} = \mathrm{c}_1\mathrm{b}^{(\mathrm{n}-1)} + \mathrm{c}_2\mathrm{b}^{(\mathrm{n}-1)} + \ldots + \mathrm{c}_{(\mathrm{n}-1)}\mathrm{b} + \mathrm{c}_\mathrm{n} \quad (1)$$

H = nilai hash
c = ASCII character value
n = n−*gram* value
b = prime number base

e-ISSN: xxxx-xxxx
p-ISSN: xxxx-xxxx

Jaccard

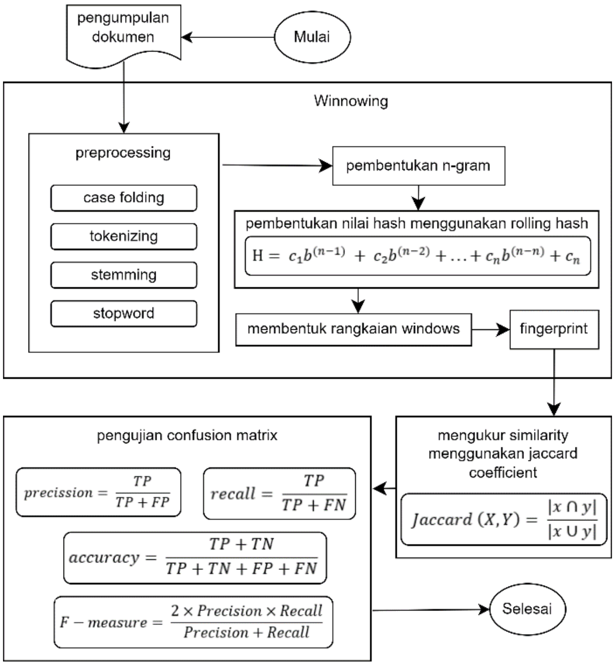
The Jaccard Coefficient or Jaccard Similarity is an algorithm developed by Paul Jaccard in 1901. This method serves to evaluate the similarities between two data sets, such as comparing one document to another based on the words used. Typically, the Jaccard method is used to compare documents and calculate the similarity of two document objects [6]. The Jaccard Coefficient can be formulated as follows:

$$\text{Jaccard}\ (X, Y) = \frac{|x\cap y|}{|x\cup y|} \quad (2)$$ gram value n−gram value

X = Document Fingerprint 1
Y = Document Fingerprint 2

The research procedure carried out can be seen in figure 2:

Figure 2. Research Procedure


pengumpulan dokumen
Mulai
Winnowing
preprocessing
case folding
tokenizing
stemming
stopword
pembentukan n-gram
pembentukan nilai hash menggunakan rolling hash
$H = c_1 b^{(n-1)} + c_2 b^{(n-2)} + \ldots + c_n b^{(n-n)} + c_n$
membentuk rangkaian windows
fingerprint
mengukur similarity menggunakan jaccard coefficient
$Jaccard\ (X,Y) = \frac{|x \cap y|}{|x \cup y|}$
pengujian confusion matrix
$precission = \frac{TP}{TP+FP}$
$recall = \frac{TP}{TP+FN}$
$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$
Selesai

1. Document collection: The first stage is to collect documents that will be used as training and test documents. This document is sourced from the repository of Mercu Buana University.
2. Preprocessing: In this stage, the documents that have been collected will go through a series of pre-processing processes. This includes converting all letters to lowercase letters (case folding), breaking the text into tokens (tokenizing), removing non-descriptive words (stopwords), and applying stemming to convert words into their basic form.
3. Formation of n-grams: At this stage, n-grams will be formed from pre-processed text. N-grams are used to describe text in a more structured form.
4. Hash value formation using rolling hash: Rolling hash is a method to generate a hash value from n-grams. This hash value is used to compare n-grams among different documents and identify possible similarities.

e-ISSN: xxxx-xxxx
p-ISSN: xxxx-xxxx

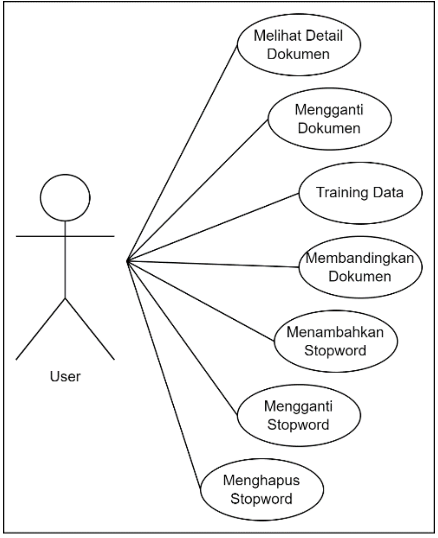
5. Creating a network of windows: This stage is the creation of a series of windows (winnowing) by selecting the smallest hash value of each window that moves through the document. This smallest hash value is considered the fingerprint of the window.
6. Fingerprint: After going through the winnowing process, a fingerprint is formed that uniquely represents the text in the window.
7. Measuring similarity using the jaccard coefficient: At this stage the fingerprint will measure the degree of similarity between the two documents being compared. This generates a similarity value that can be used to determine the extent to which two documents are similar.
8. Confusion matrix testing: In the last stage, similarity measurement results are used to perform tests using confusion matrix. This includes precision, recall, accuracy, and F-measurement measurements to evaluate how well the system can detect plagiarism in the documents being compared.

RESULT AND DISCUSSION

A. Use Care

Use case diagrams are used to model the functionality of a system or software from the user's point of view [9]. Users can view document details, replace documents, train data, compare documents, add stopwords, replace stopwords and delete stopwords can be seen in figure 3:

Figure. 3 Use Case Diagram



B. Dataset

The dataset used in this study is chapter 1 of the final project of Mercu Buana University students obtained from https://repository.mercubuana.ac.id. The number of documents used for this study is 10 training documents and 1 test document for comparison.

C. Preprocessing

The preprocessing stage is carried out to clean the data and transform the data into a more structured [10] preprocessing dilakukan untuk membersihkan data dan mentransformasikan data menjadi lebih terstruktur. At this stage, the data is processed into a more appropriate format and easy to process by the algorithm or model to be used [10]1]. In this study, there are 4 stages of preprocessing , namely, case folding, tokenizing, stemming, and stopword.

D. N-Gram

N-Gram is the process of dividing a number of characters or words from a group of terms. The N-Gram method is used to generate characters or words. The N-Gram method is used to retrieve the letters of several characters from a word in a continuous manner until the end of the document [13].

The process of forming n-grams in texts that have gone through the preprocessing stage is carried out by determining the desired length of n-grams, in this study n=7, which means that the text will form a sequence of seven words or characters. Then, the text will move one step forward each time to form the next n-gram. Here are the steps of n-gram formation:

1. Take the first seven characters of the text:
   "industr "
   Save this n-gram as the first n-gram.
2. Slide one character forward:
   Take the 2nd character to the 8th character of the text: "ndustri"
   Save this n-gram as the second n-gram.

Continue this process until you reach the end of the text. Each time, advance one character and take the next seven characters to form different n-grams. So that the n-grams formed from the text:

First N-gram: "industr "
Second N-gram: "ndustri"
Third N-gram: "dustrii"
And so on until the end of the text the result of which can be seen in table 1:

Table 1. Formation Of N-Gram

N-Gram = 7
industr ndustri dustrii ustriin striint triinte riinter iintern interne nternet ternetm ernetma rnetmai netmain etmaing tmainga maingam aingame ingameb ngamebe gamebel amebela mebelan ebelanj belanja elanjak lanjako anjakom njakomu jakomun akomuni komunik omunika munikas unikasi nikasia ikasiaj kasiaja asiajar siajaro iajaror ajarora jaroran arorang roranga orangak rangaks angakse ngakses gaksesi aksesin ksesint sesinte esinter sintern interne nternet terneta ernetan rnetana netanak etanaka tanakan anakana nakanak akanaka kanakaj anakaja nakajar akajaro kajaror ajarora jaroran arorang rorangd orangde rangdew angdewa ngdewas gdewasa dewasao ewasaor wasaora asaoran saorang aorangt orangtu rangtua angtuab ngtuabu gtuabuk tuabuka uabukaw abukawe bukaweb ukawebs kawebsi awebsit website ebsites bsitesa sitesal itesala tesalah

esalahc salahco alahcon lahcont ahconto hcontoh

E. Rolling Hash

In this stage, any n-grams that have been formed will be converted into a hash value using the rolling hash method. To calculate the hash value of the word "industry" with values b=3 and n=7.

Taking the ASCII value of each character from the word "industry" is obtained the ASCII value as follows:

i = 105
n = 110
d = 100
u = 117
s = 115
t = 116
r = 114

Thus, the results can be calculated based on equation 1 as follows:

$$H = 105 \times 3^6 + 110 \times 3^5 + 100 \times 3^4 + 117 \times 3^3 115 \times 3^2 + 116 \times 3 + 114$$

$$H = (105 \times 729) + (110 \times 243) + (100 \times 81) + (117 \times 27) + (115 \times 9) + (116 \times 3) + 114$$

$$H = 76545 + 26730 + 8100 + 3159 + 1035 + 348 + 114$$

$$H = 116031$$

So, the hash value of the word "industr" with values b=3 and k=7 are 116031. If all hash values are calculated, the result can be seen in table 2:

Table 2. Rolling Hash Results

industr: 116031	ingameb: 115619	ikasiaj: 114850
ndustri: 118563	ngamebe: 117323	kasiaja: 115012
dustrii: 115224	gamebel: 111507	asiajar: 111141
ustriin: 127082	amebela: 109357	siajaro: 121395
striint: 125483	mebelan: 116042	iajaror: 112794
triinte: 125045	ebelanj: 109849	ajarora: 108844
riinter: 121557	belanja: 108757	jaroran: 114503
iintern: 115463	elanjak: 112052	arorang: 111790
interne: 116855	lanjako: 115380	roranga: 123328
nternet: 121046	anjakom: 110053	orangak: 120773
ternetm: 122677	njakomu: 118137	rangaks: 119677

ernetma:	jakomun:	angakse:
114436	113951	109814
rnetmai:	akomuni:	ngakses:
122526	110136	117418
netmain:	komunik:	gaksesi:
118370	118376	111789
etmaing:	omunika:	aksesin:
114643	121216	110216
tmainga:	munikas:	ksesint:
123139	121006	118625
maingam:	unikasi:	sesinte:
115834	124740	121967
aingame:	nikasia:	…
109220	118438	

F. Winnowing

This stage is the creation of a series of windows (winnowing) by selecting the smallest hash value of each window that moves through the document. This smallest hash value is considered the fingerprint of the window [16]. In this study, window length = 5 is used, the visualization of which can be seen in table 3:

Table 3. Winnowing Results

industr:	triinte:	terneta:
116031	125045	122665
ndustri:	riinter:	ernetan:
118563	121557	114413
dustrii:	iintern:	rnetana:
115224	115463	122449
ustriin:	interne:	netanak:
127082	116855	118136
striint:	nternet:	etanaka:
125483	121046	113935
ndustri:	riinter:	ernetan:
118563	121557	114413
dustrii:	iintern:	rnetana:
115224	115463	122449
ustriin:	interne:	netanak:
127082	116855	118136
striint:	nternet:	etanaka:
125483	121046	113935
triinte:	terneta:	tanakan:
125045	122665	121028
dustrii:	iintern:	rnetana:
115224	115463	122449
ustriin:	interne:	netanak:
127082	116855	118136
striint:	nternet:	etanaka:
125483	121046	113935

triinte: 125045	terneta: 122665	tanakan: 121028
riinter: 121557	ernetan: 114413	anakana: 109489
ustriin: 127082	interne: 116855	netanak: 118136
striint: 125483	nternet: 121046	etanaka: 113935
triinte: 125045	terneta: 122665	tanakan: 121028
riinter: 121557	ernetan: 114413	anakana: 109489
iintern: 115463	rnetana: 122449	nakanak: 116435
striint: 125483	nternet: 121046	etanaka: 113935
triinte: 125045	terneta: 122665	tanakan: 121028
riinter: 121557	ernetan: 114413	anakana: 109489
iintern: 115463	rnetana: 122449	nakanak: 116435
interne: 116855	netanak: 118136	akanaka: 108832
		…

After the window series is formed, the next step is to take the smallest hash value of each window to be used as a fingerprint, the final result can be seen in table 4:

Table 4. Fingerprint Document 1

dustrii: 115224	asiajar: 111141	angdewa: 109849
iintern: 115463	ajarora: 108844	dewasao: 111138
ernetma: 114436	arorang: 111790	asaoran: 110939
etmaing: 114643	angakse: 109814	angtuab: 110360
aingame: 109220	aksesin: 110216	abukawe: 108224
amebela: 109357	sesinte: 114510	awebsit: 111923
belanja: 108757	terneta: 114413	ebsites: 111055
anjakom: 110053	etanaka: 113935	bsitesa: 112375

akomuni: 110136	anakana: 109489	esalahc: 113631
komunik: 118376	akanaka: 108832	alahcon: 108956
ikasiaj: 114850	akajaro: 108759	ahconto: 108450

As a comparison document, the same process is also carried out on the second document so that the fingerprint of the second document can be seen in table 5:

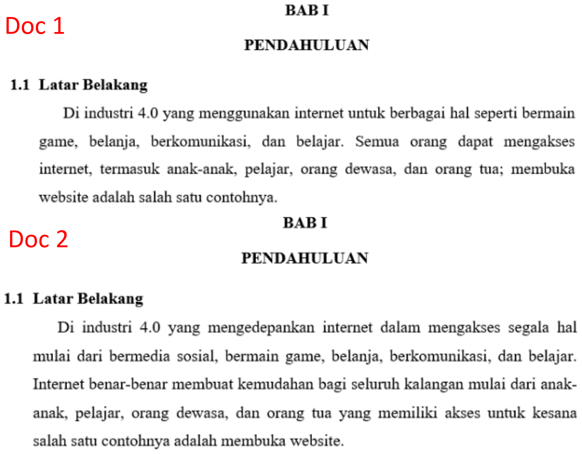
Table 5. Fingerprint Document 2

dustrii: 115224	akomuni: 110136	ajarora: 108844
iintern: 115463	komunik: 118376	arorang: 111790
ernetak: 114410	omunika: 114850	angdewa: 109849
etakses: 114016	asiajar: 111141	dewasao: 111138
aksesse: 110237	ajarint: 108797	asaoran: 110939
essegal: 114942	arinter: 111351	angtuam: 110371
egalamu: 110748	ernetmu: 114456	amilika: 109984
alamula: 109231	etmudah: 115100	kaksesk: 114707
amulaim: 110890	dahselu: 109350	akseske: 110213
aimedia: 109096	ahselur: 109464	eskesan: 114404
ediasos: 110536	eluruhk: 113900	esanasa: 113716
diasosi: 110826	hkalang: 113896	anasala: 109699
asosial: 112098	alangmu: 109155	alahcon: 108956
almaing: 109783	angmula: 110203	ahconto: 108450
aingame: 109220	laianak: 114815	contohb: 112595
amebela: 109357	aianaka: 108346	hbukawe: 113327
belanja: 108757	akanaka: 108832	bukaweb: 112631
anjakom: 110053	akajaro: 108759	awebsit: 111923

G. Jaccard Coefficient

At this stage, the fingerprint will be measured by the level of similarity between two documents compared using the Jaccard Coefficient. The contents of documents 1 and 2 can be seen in figure 4:

Figure 4. Contents of Documents 1 and 2



Doc 1

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di industri 4.0 yang menggunakan internet untuk berbagai hal seperti bermain game, belanja, berkomunikasi, dan belajar. Semua orang dapat mengakses internet, termasuk anak-anak, pelajar, orang dewasa, dan orang tua; membuka website adalah salah satu contohnya.

Doc 2

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di industri 4.0 yang mengedepankan internet dalam mengakses segala hal mulai dari bermedia sosial, bermain game, belanja, berkomunikasi, dan belajar. Internet benar-benar membuat kemudahan bagi seluruh kalangan mulai dari anak-anak, pelajar, orang dewasa, dan orang tua yang memiliki akses untuk kesana salah satu contohnya adalah membuka website.

The Jaccard Coefficient is used to measure the similarity between two sets, which in this case represents two documents [17]. Thus, the results can be calculated based on equation 2 with the fingerprint of document 1 (X) obtained from table VIII and the fingerprint of document 2 obtained from table 5 (Y).

X = {115224, 115463, 114436, 114643, 109220, 109357, 108757, 110053, 110136, 118376, 114850, 111141, 108844, 111790, 109814, 110216, 114510, 114413, 113935, 109489, 108832, 108759, 109849, 111138, 110939, 110360, 108224, 111923, 111055, 112375, 113631, 108956, 108450}

Y = {115224, 115463, 114410, 114016, 110237, 114942, 110748, 109231, 110890, 109096, 110536, 110826, 112098, 109783, 109220, 109357, 108757, 110053, 110136, 118376, 114850, 111141, 108797, 111351, 114456, 115100, 109350, 109464, 113900, 113896, 109155, 110203, 114815, 108346, 108832, 108759, 108844, 111790, 109849, 111138, 110939, 110371, 109984, 114707, 110213, 114404, 113716, 109699, 108956, 108450, 112595, 113327, 112631, 111923}

$X \cap Y$ = {115224, 115463, 109220, 109357, 108757, 110053, 110136, 118376, 114850, 111141, 108844, 111790, 109814, 109849, 111138, 110939, 108832, 108759, 108956, 108450}
= 20

$X \cup Y$ = {114436, 115463, 113935, 114707, 110360, 115224, 109849, 114456, 108832, 111138, 110371, 111141, 109350, 109096, 110890, 108844, 109357, 111923, 113716, 110136, 108346, 114510, 110939, 114016, 109155, 118376, 111351, 110203, 114815, 109699, 110213, 110216, 112631, 109464, 108956, 115100, 110748, 110237, 109984, 114850, 108450, 109220, 111790, 109231, 113327, 109489, 108224, 110536, 111055, 114643, 112595, 108757, 108759, 109783, 113631, 112098, 114404, 110053, 113896, 114410, 110826, 113900, 114413, 109814, 112375, 108797, 114942}

= 67

then it will generate the value:

$$\text{Jaccard}\ (X, Y) = \frac{|20|}{|67|} = 0{,}29851 \text{ x } 100\% = 29{,}851\%$$

So, the level of similarity between document 1 and document 2 after calculating using the jaccard coefficient method is 29.851%.

H. Confusion Matrix Testing

The level of similarity obtained using the jaccard coefficient method will go through an accuracy test stage using the Confusion Matrix. The Confusion Matrix is a table used to describe the predictions of the classification algorithm against the actual values in the positive and negative categories [18]. This testing stage is aimed at measuring the accuracy of the results obtained from the system and with a plagiarism threshold of 15%. The results of the comparison between the test document and the training document obtained from chapter 1 in the student's final project can be seen in table 6.

Table 6. Hash Rolling Results

No	Test Documents	Training Documents	Slices (∩)	Combined (∪)	Similarities
1	doc1.docx	doc1.docx	539	539	100 %
2		doc2.docx	82	1727	4.74%
3		doc3.docx	68	1046	6.50%
4		doc4.docx	84	1295	6.48%
5		doc5.docx	62	1175	5.27%
6		doc6.docx	75	1064	7.04%
7		doc7.docx	49	930	5.26%
8		doc8.docx	59	911	6.47%
9		doc9.docx	86	1304	6.59%
10		doc10.docx	67	1151	5.82%

From the results of the comparison in table X, then classification is carried out with a confusion matrix whose results can be seen in table 7.

Table 7. Ilustration of Classification with Confusion Matrix

No	Test Documents	Training Documents	Jaccard		Treshold 15%		Classification
			+	-	+	-	
1	doc1.docx	doc1.docx	✓		✓		TP
2		doc2.docx		✓		✓	TN
3		doc3.docx		✓		✓	TN
4		doc4.docx		✓		✓	TN
5		doc5.docx		✓		✓	TN
6		doc6.docx		✓		✓	TN
7		doc7.docx		✓		✓	TN
8		doc8.docx		✓		✓	TN
9		doc9.docx		✓		✓	TN
10		doc10.docx		✓		✓	TN

11	doc2.docx	doc1.docx+ doc2.docx	✓	✓	TP
12	doc4.docx	doc3.docx+ doc4.docx	✓	✓	TP
13	doc6.docx	doc5.docx+ doc6.docx	✓	✓	TP
14	doc8.docx	doc7.docx+ doc8.docx	✓	✓	TP
15	doc10.docx	doc9.docx+ doc10docx	✓	✓	TP

1. Precision

$$\text{precission} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \ (3)$$

$$\text{precission} = \frac{6}{6+0} \times 100\%$$

$$\text{precission} = 100\%$$

2. Recall

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \ (4)$$

$$\text{recall} = \frac{6}{6+0} \times 100\%$$

$$\text{recall} = 100\%$$

3. Accuracy

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \ (5)$$

$$\text{accuracy} = \frac{6+9}{6+9+0+0}$$

$$\text{accuracy} = 100\%$$

4. F-measurement

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

$$\text{F} - \text{measure} = \frac{2 \times 100 \times 100}{100 + 100}$$

$$\text{F} - \text{measure} = 1$$

Based on the results of accuracy testing using the Confusion Matrix, the system successfully classified documents with a level of precision, recall, and accuracy of 100%. These results show the system's ability to accurately identify documents that have a level of similarity above the set plagiarism limit, which is 15%. Analysis using precision, recall, accuracy, and F-measurement shows that the system achieves the optimal balance between precision and sensitivity.

I. Stopword Testing

Stopword effectiveness testing was conducted to evaluate the impact of the use of word elimination techniques on system accuracy. In this test, there are three different conditions:

1. Without using a stopword
2. Use 70 stopwords

3. Use 140 stopwords

First of all, the system is run without the use of stopwords, so the entire word in the text is considered important. The results can be seen in table 8.

Table 8. Interrupted Results

No	Test Documents	Training Documents	Slices (∩)	Combined (∪)	Similarities
1	doc1.docx	doc1.docx	737	737	100%
2		doc2.docx	202	2301	8.77%
3		doc3.docx	124	1374	9.02%
4		doc4.docx	135	1805	7.47%
5		doc5.docx	124	1590	7.79%
6		doc6.docx	106	1467	7.22%
7		doc7.docx	105	1273	8.24%
8		doc8.docx	91	1292	7.04%
9		doc9.docx	145	1804	8.03%
10		doc10.docx	116	1523	7.61%

Then, the test was carried out using 70 stopwords. The use of a smaller number of stopwords is intended to see if this reduction has any effect on system accuracy. The results can be seen in table 9.

Table 9. 70 Stopword Results

No	Test Documents	Training Documents	Slices (∩)	Combined (∪)	Similarities
1	doc1.docx	doc1.docx	622	622	100%
2		doc2.docx	131	2001	6.54%
3		doc3.docx	85	1202	7.07%
4		doc4.docx	100	1526	6.55%
5		doc5.docx	89	1368	6.50%
6		doc6.docx	91	1236	7.36%
7		doc7.docx	72	1078	6.67%
8		doc8.docx	75	1081	6.93%
9		doc9.docx	112	1509	7.42%
10		doc10.docx	88	1343	6.55%

The last test was done by increasing the number of stopwords to 140. The purpose of this test was to evaluate whether the addition of a stopword would provide an improvement in the accuracy of the plagiarism detection system. The results can be seen in table 10:

e-ISSN: xxxx-xxxx
p-ISSN: xxxx-xxxx

Table 10. 140 Stopwords Resutls

No	Test Documents	Training Documents	Slices (∩)	Combined (∪)	Similarities
1	doc1.docx	doc1.docx	539	539	100 %
2		doc2.docx	82	1727	4.74%
3		doc3.docx	68	1046	6.50%
4		doc4.docx	84	1295	6.48%
5		doc5.docx	62	1175	5.27%
6		doc6.docx	75	1064	7.04%
7		doc7.docx	49	930	5.26%
8		doc8.docx	59	911	6.47%
9		doc9.docx	86	1304	6.59%
10		doc10.docx	67	1151	5.82%

The test results show that the application of the stopword technique can improve the accuracy of the system. The more stopwords used, the more effective the system is at measuring the level of similarity. This indicates that non-descriptive words that are not omitted can increase the level of similarity between documents. By eliminating more stopwords, the system can be more effective in distinguishing documents.

J. N-Gram Testing

N-gram testing is performed to determine the optimal amount of n-grams so as to achieve a balance between accuracy and sensitivity. In this test, a comparison of system performance was carried out using different n-gram values, namely 5, 7, and 9. The test results can be seen in Table 1:

Table 11. Illustration of Classification with Confusion Matrix

No	Test Documents	Training Documents	Similarity Levels		
			N=5	N=7	N=9
1	doc1.docx	doc1.docx	100 %	100 %	100 %
2		doc2.docx	26.96%	4.74%	1.45%
3		doc3.docx	24.08%	6.50%	3.52%
4		doc4.docx	25.09%	6.48%	2.07%
5		doc5.docx	25.17%	5.27%	1.68%
6		doc6.docx	25.45%	7.04%	2.50%
7		doc7.docx	24.78%	5.26%	2.66%
8		doc8.docx	23.87%	6.47%	3.47%
9		doc9.docx	27.89%	6.59%	1.94%
10		doc10.docx	24.81%	5.82%	1.69%

The results showed that n=7 resulted in an optimal level of similarity in achieving the similarity between accuracy and sensitivity. This indicates that a model with a value of n=7 is effective in detecting plagiarism in complex and long texts, providing accurate results without sacrificing sensitivity to word variation.

CONCLUSION

Based on the results of the study, it can be concluded that:

1. The similarity index detection system uses the Winnowing algorithm with the Jaccard Coefficient approach with n-grams of 7 to achieve optimal results with 100% precision, recall, and accuracy results.
2. The application of the word elimination technique (stopword) is able to increase the accuracy of the system.
3. The degree of similarity between documents can be effectively measured, and the performance of the Winnowing algorithm can be well evaluated using the Jaccard Coefficient method.
4. If the n-gram value entered is smaller, then the level of similarity between documents will result in a high value, and vice versa.

REFERENCE

[1] M. A. Shadiqi, “Memahami dan Mencegah Perilaku Plagiarisme dalam Menulis Karya Ilmiah,” *Buletin Psikologi*, vol. 27, no. 1, p. 30, Jun. 2019, doi: 10.22146/buletinpsikologi.43058.

[2] N. Nurdin, R. Rizal, and R. Rizwan, “Pendeteksian Dokumen Plagiarisme dengan Menggunakan Metode Weight Tree,” *Telematika*, vol. 12, no. 1, p. 31, Feb. 2019, doi: 10.35671/telematika.v12i1.775.

[3] S. Sariffuddin, K. D. Astuti, and R. Arthur, “Investigating Plagiarism: The Form and The Motivation in Performing Plagiarism in High Education,” *Journal of Education and Learning (EduLearn)*, vol. 11, no. 2, pp. 172–178, May 2017, doi: 10.11591/edulearn.v11i2.5994.

[4] D. D. Sinaga and S. Hansun, “Indonesian text document similarity detection system using rabin-karp and confix-stripping algorithms,” *International Journal of Innovative Computing, Information and Control*, vol. 14, no. 5, pp. 1893–1903, Oct. 2018, doi: 10.24507/ijicic.14.05.1893.

[5] S. Sugiono, H. Herwin, H. Hamdani, and E. Erlin, “Aplikasi Pendeteksi Tingkat Kesamaan Dokumen Teks: Algoritma Rabin Karp Vs. Winnowing,” *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 82–93, May 2018, doi: 10.31849/digitalzone.v9i1.1242.

[6] S. Sunardi, A. Yudhana, and I. A. Mukaromah, “Implementasi Deteksi Plagiarisme Menggunakan Metode N-Gram Dan Jaccard Similarity Terhadap Algoritma Winnowing,” *Transmisi: Jurnal Ilmiah Teknik Elektro*, vol. 20, no. 3, pp. 105–110, 2018.

[7] Y. Nurdiansyah, F. Nur Muharrom, and F. Firdaus, “Implementation of Winnowing Algorithm Based K-Gram to Identify Plagiarism on File Text-Based Document,” in *MATEC Web of Conferences*, EDP Sciences, Apr. 2018. doi: 10.1051/matecconf/201816401048.

[8] E. G. , W. A. , & H. S. Hasan, “The Implementation of Winnowing Algorithm for Plagiarism Detection in Moodle-based E-learning,” 2018.

[9] Wawan Gunawan and Bagus Seno Prasetyo Diwiryo, “Implementasi Algoritma Fuzzy C-Means Clustering Sistem Crowdfunding pada SektorIndustri Kreatif Berbasis Web,” *JEPIN (Jurnal Edukasi dan Penelitian Informatika)* , vol. 6, 2020.

[10] Ragil Dimas Himawan and Eliyani Eliyani, “Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi,” *JEPIN (Jurnal Edukasi dan Penelitian Informatika)* , vol. 7, 2021.

[11] D. Soyusiawaty and F. Rahmawanto, “Similarity Detector on the Student Assignment Document Using Levenshtein Distance Method,” in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Nov. 2018, pp. 656–661. doi: 10.1109/ISRITI.2018.8864339.

[12] E. Y. Puspaningrum, B. Nugroho, A. Setiawan, and N. Hariyanti, “Detection of Text Similarity for Indication Plagiarism Using Winnowing Algorithm Based K-gram and Jaccard Coefficient,” *J Phys Conf Ser*, vol. 1569, no. 2, p. 022044, Jul. 2020, doi: 10.1088/1742-6596/1569/2/022044.

[13] W. G. S. Parwita, I. G. A. A. D. Indradewi, and I. N. S. W. Wijaya, “String Matching based Plagiarism Detection for Document in Bahasa Indonesia,” in *2019 5th International Conference on New Media Studies (CONMEDIA)*, IEEE, Oct. 2019, pp. 54–58. doi: 10.1109/CONMEDIA46929.2019.8981821.

[14] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, “Sentiment analysis in the sales review of Indonesian marketplace by utilizing Support Vector Machine,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 4, no. 1, pp. 57–64, 2018.

[15] P. Buttar, J. Kaur, and P. Kaur Buttar, “A Systematic Review on Stopword Removal Algorithms,” 2018, [Online]. Available: http://www.ijfrcsce.org

[16] Y. Arifin, S. M. Isa, L. A. Wulandhari, and E. Abdurachman, “Plagiarism Detection for Indonesian Language using Winnowing with Parallel Processing,” *J Phys Conf Ser*, vol. 978, p. 012082, Mar. 2018, doi: 10.1088/1742-6596/978/1/012082.

[17] K. Yang and Y. Xu, “An effective method for complex network community detection based on hierarchical splitting,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2019, pp. 10–14. doi: 10.1145/3325730.3325747.

[18] D. Valero-Carreras, J. Alcaraz, and M. Landete, “Comparing two SVM models through different metrics based on the confusion matrix,” *Comput Oper Res*, vol. 152, p. 106131, Apr. 2023, doi: 10.1016/j.cor.2022.106131.