Bit-Tech (Binary Digital - Technology)
Vol.7, No.2, December 2024
Available online at: http://jurnal.kdi.or.id/index.php/bt

Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods

Jimmy Halim[1)*], Desiyanna Lasut[2)]

[1)2)]*Buddhi Dharma University*
Imam bonjol, Tangerang, Indonesia
[1)]jimmyhalim345@gmail.com
[2)]desiyanna.lasut@buddhidharma.ac.id

Article history:

Received 01 Oct 2024;
Revised 29 Oct 20xx;
Accepted 05 Nov 20xx;
Available online 27 Dec 2024

Keywords:

Cosine Similarity
Detection
Plagiarism
TF-IDF
Website

Abstract

In the era of technology and information which is developing very rapidly recently, this has resulted in easy access to information which makes the learning process easier in the world of education, but this ease also triggers acts of plagiarism which is a serious threat to science. Plagiarism is an act of stealing or taking someone else's work without giving proper attribution or you could say without citing that person. Therefore, an application was developed that can overcome this problem, namely a plagiarism detection application that uses the TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity algorithm methods. TF-IDF and Cosine Similarity will be implemented into the application to carry out the calculation process which will ultimately provide results in the form of a percentage of the calculations that have been carried out. This plagiarism application is designed to detect similarities between documents in the database and user documents. The processes that occur in the application include preprocessing processes, tf-idf calculations, and cosine similarity calculations. The results of the tests carried out can be said to be consistent because the results of manual and application tests show percentage results of 4% and 4.34%. The application will also be website-based, and will be designed in such a way that it can be used to detect plagiarism.

I. INTRODUCTION

In the era of technology and information which is developing very rapidly recently, anyone can easily search and find the information they need. The impact of this development is very large, especially in the world of education. College students now have wide access to various sources of information such as books on the internet, journals on Google, and tutorials on YouTube. In the world of education which is supported by technology, information from e-books, articles and journals both international and national with varying reputations can be easily accessed via the internet [1]. Although this convenience provides many benefits, there are also negative impacts that arise, such as increasing acts of plagiarism.

Plagiarism is an unlawful act that poses a serious threat to the world of science. Perpetrators of plagiarism usually steal other authors' works into their work without citing the original references [2]. All these organizations agree that violations in scientific research can occur basically through three practices condemned by researchers: Falsifying research data, Falsifying results, and authorship fraud, which means the undue appropriation of content belonging to another author. without proper credit attribution. In addition, other condemnable practices such as redundancy in publication or self-plagiarism are also considered in the same category as careless handling of research subjects or piracy [3].

Plagiarism can be defined as the act of taking over information, data and knowledge that is actually the work of another person without citing the original source. To overcome this problem, plagiarism detection tools were developed. This tool really supports the world of education because it can maintain academic fairness. Plagiarism often occurs in student assignments, and this action is very detrimental to the world of education.

The Directorate General of Higher Education on January 4 2012 recorded 21 universities, some of which are leading universities in Indonesia. Cases that have shocked the world of education in the country have been reported, among others case of former rector of the State Islamic University (UIN) Maliki Malang Prof. Mudjia

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi: 10.32877/bt.v7i2.1697

Rahardjo is suspected plagiarized eight papers written by guidance students in a book entitled "Sosiolinguistik Qurani". This book was published by UIN press in 2008. Then, the case former Chancellor of Jakarta State University (UNJ) Prof. Djaali who was dismissed by the Minister of Research Technology and Higher Education for committing massive plagiarism cases [4].

Another case is Chancellor of Halu Oleo University (UHO) Dr. Muhammad Zamrun who was suspected of plagiarism by 30 UHO professors. Zamrun is strongly suspected of plagiarizing a number of scientific papers in journals he wrote. In another case, the Chancellor of Sultan Ageng Tirtayasa University, Banten, Prof. Sholeh Hidayat moment with the rank of Intermediate Principal Supervisor/IVd received a verbal warning for violating written copyright. Lecturer at UIN Sunan Gunung Djati Bandung, Ade Juhana allegedly completed his dissertation with hijacked Prof.'s thesis. Tihami and Mohamad Hudaeri's book. Then, lecturer at the Bogor Agricultural Institute, Heri Ahmad Sukria was involved in allegations of plagiarism because of a book entitled ""Sumber dan Ketersediaan Bahan Baku Pakan di Indonesia [4]".

The various tools used to detect plagiarism are usually web applications available online, such as Turnitin, which has been used by many lecturers and teachers. Therefore, this research aims to create a web-based application that can detect plagiarism. This application uses a string matching algorithm in text documents to search for similar words between documents. The algorithms used are TF-IDF (Term Frequency-Inverse Document Frequency) and Cosine Similarity. By matching strings in the compared documents, this application produces output that shows how closely the documents are similar to each other.

This technique is used after going through a lot of previous research with articles that have been read, this technique is used because this technique has a thorough calculation method for each word carried out by the TF-IDF technique and the results of the TF-IDF will be recalculated by the Cosine technique Similarity, this causes detailed calculations without missing a single word of calculation, this is what makes a significant difference from other techniques. In previous research, string matching algorithms were applied to find similarities between documents, but were not accompanied by manual calculations that showed the TF-IDF and Cosine Similarity calculation processes in detail. This research is expected to provide a new contribution by displaying manual calculation steps for each algorithm stage, which have not been implemented in previous research.

It is hoped that this application can help lecturers and students reduce the level or acts of plagiarism that often occur. This aims to get originality from a student's assignment, where now it is an era where it is very easy to get information and also easily take it without any responsibility, this can also help students to do the assignment with their own hard work. Likewise, lecturers can assess student assignments efficiently, in this way lecturers will be able to grade student assignments fairly.

II. LITERATURE REVIEW

Plagiarism is the process of copying and pasting other people's intellectual products which are misused without mentioning the names of the original authors, inventors and initiators [2], or Plagiarism can be defined as signing or presenting oneself as the author of another person's artistic or scientific work. or you could say copying other people's work [3]. Plagiarism can be divided into 4 categories, and each category has different ways of plagiarism. You can see the 4 categories, namely:

A. Verbatim Plagiarism

In Verbatim Plagiarism, the author directly takes someone else's work or opinion without changing it, it can be said to be exactly the same. Verbatim Plagiarism is often known as Copy-Paste.

B. Patchwork Plagiarism

Patchwork Plagiarism is a combination type of plagiarism. Because the writer will combine pieces of other people's ideas or work which are finally recognized as their own ideas or concepts. Perpetrators of Patchwork Plagiarism often combine pieces of ideas or sentences from several authors to make one complete sentence without citing the source at all.

C. Paraphrase Plagiarism

Paraphrase Plagiarism is usually done by composing sentences from original words, but claiming that these words are your own writing or ideas. Writing from original words is often changed in such a way as to avoid plagiarism. Perpetrators of Paraphrasing Plagiarism are usually dishonest and do not want to include sources.

D. Plagiarism of key word or phrases

The perpetrator who uses word or key phrase plagiarism will take key words or phrases from an article. The writings taken will be compiled, then made into new writing by the perpetrator. The perpetrator of this plagiarism of key words or phrases will describe and assemble the writings that have been taken into a complete sentence to form a new meaning without mentioning the source at all.

Plagiarism is one of the enemies in the world of education that is very difficult to eradicate, this is because many perpetrators do not think about the consequences of doing such immoral things. This causes great losses for the victims whose research results are used as they please. without including the victim's name. Therefore, a tool is needed that can be used to detect this plagiarism.

Plagiarism detection is one of the important aspects in the academic world to maintain the originality of scientific work. TF-IDF (Term Frequency-Inverse Document Frequency) combined with Cosine Similarity is a technique widely used in plagiarism detection. The TF-IDF method is a method for calculating the weight of a word (term) in the document. This method too is known to be efficient, easy and has accurate results . This method combines two concepts for calculations weight, that is, the frequency with which a word appears in it a particular document and the inverse frequency of that document contains the word [5].

From previous studies using the Rabin-Karp algorithm, this method is not as detailed as the TF-IDF and Cosine Similarity methods. The Rabin-Karp algorithm is a word search algorithm that searches for patterns in the form of substrings in a text using a hashing function [6]. while TF-IDF and Cosine Similarity calculate the weight of a word (term) in the document first and then carry out the Cosine Similarity calculation process.

This research aims to develop a web-based plagiarism detection application. In this research, the researchers will use several methods to create a web-based plagiarism detection application, namely: the Waterfall method, preprocessing method, Entity Relationship Diagram (ERD) method, Data Flow Diagram (DFD) method, black box testing method, TF-IDF method, Cosine Similarity method and, text mining method.

III. METHODS

Data is information, or information in the form of numbers, and can also be in the form of categories resulting from a process of observing, calculating, and measuring a variable that describes a problem [7]. Data is very important because it provides the information needed to identify patterns and relationships between relevant variables. By analyzing data thoroughly, we can obtain insights that may be needed for further analysis. The data obtained will be processed properly so that it can be easily utilized. Data can also be grouped into two types, namely:

1. **Primary Data**
   Primary data is data obtained directly from organizations, institutions, bodies, or individuals directly from the source [7]. This data is usually gathered through interviews, observations, or surveys. Once collected, this primary data will be compiled and used for the subsequent research process.
2. **Secondary Data**
   Secondary data is data needed by an organization or company but collected by another party where the data is already processed and finished [7]. Examples of secondary data include population statistics, journal articles, and tabular data.

Data can be extracted to become useful information. In this research, the data used is extracted from journals and collected to become test data. This process can be considered a text mining process. Text Mining is the process of examining large amounts of documents to discover new information or to answer specific research questions by identifying key facts, relationships, and statements. Once extracted, this information is structured into a format that can be further analyzed or directly presented using grouped HTML tables, mind maps, charts, etc. [8]. Data will be entered via the file upload menu which has been designed to accept PDF files.

The text mining process is one of the most important processes in this research because it forms the core of a plagiarism detection website. In this application, researchers create their own database to store the retrieved information. The database will be formed using the MySQL database with the help of the XAMPP application. Xampp is an application that is formed in such a way that it plays the role of a local web server, this means that xampp will provide a localhost which will act as a web server and also a database system [9]. Data that will be used in this research is data extracted from Indonesian language journals which are collected to become test data and user data which will be used as training data.

Xampp is also an application in which several software collections are available such as PHPMyAdmin, Apache, MySQL, PHP, Perl, and others. This makes it easier to use because there is no need to install the software one by one[10]. This database will be the basic basis for storing data that will be sent by the application later and the plagiarism detection website application will be created using several methods, namely

A. Waterfall Method

Waterfall is a structured and sequential approach that allows implementation to remain focused on each development stage with full attention, from the initial planning stage to the implementation stage [11]. It can be interpreted that the waterfall method is an application development method that is based on certain sequences. The waterfall sequence has several steps as in Fig. 1 [12]:

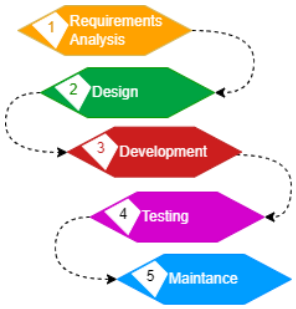

Fig. 1 Waterfall

1) *Requirements Analysis*
In the first stage there is a process of analyzing the application towards the needs of the user, this process will be carried out by analyzing websites that have the same work system as this research, this analysis will become a document of requirements that will be needed to maximize the work of the application that will be created, such as features and functions that will be developed in the application system.
2) *Design*
At this stage, the system design process occurs, such as designing the application interface, designing the database, and designing the functions that will be implemented. The results of this stage will produce a clear design to create a detailed application architecture.
3) *Development*
After the design process is successful, the process of implementing the design is carried out, these processes involve predetermined codes. This process will produce an application that runs as designed to create a website-based plagiarism application.
4) *Testing*
After the application has been developed, a testing process occurs in which the testing process will be carried out. This process aims to ensure that the application's working system functions well and runs in accordance with the specified requirements.
5) *Maintenance*
After the testing process is successful, the Maintenance process will be carried out, this is done to monitor system performance, if errors or bugs occur that were previously unknown. A repair process will be carried out, this is done to ensure that the bug no longer exists to interfere with the application's work process. Maintenance is also carried out to improve the application work system to make it more efficient.

B. Preprocessing

Preprocessing is the preparation stage before all subsequent models and algorithms are implemented; for example word segmentation for Chinese, Japanese, Vietnamese and other languages that may require word segmentation. [13]. Preprocessing will run after data from the user is received, the data must be in PDF format first so that further processing can be carried out. Preprocessing has several main techniques [14], namely:

1) *Lowercasing*
Lowercasing is carried out so that the data that will be used for research is in lower case without using capital letters. This is done because the computer will distinguish letters if the text is uppercase and lowercase. The training data that will be analyzed in this research will have all lowercase letters, so that if there is data using capital letters it will be missed or not detected by the training data that is available for use. examples can be seen in TABLE 1

TABLE 1
LOWERCASING

Before	After
Aku	aku
Jika	jika
KaMu	kamu

2) *Noise Removal*
This noise removal is carried out to remove strange symbols that exist before further analysis is carried out. examples can be seen in TABLE 2

TABLE 2
NOSIE REMOVAL

Before	After
adap$n jika s#pe*ti it#	adapun jika seperti itu
jik@ h%l Te#seb*t	jika hal tersebut

3) Stopword

Stopwords contain words that are less meaningful for analysis, this can reduce the quality of the data that will be analyzed. The purpose of eliminating stop words is to reduce the amount of text that will be processed further later, thereby causing the process to be carried out later to be faster and not take up much time. To remove or delete stop words in Indonesian, use the Python Sastrawi library.

4) Text Transformation

Text Transformation or known as Tokenization. Tokenization is the process of changing initially larger text into a collection of several small texts, this is also called a token. Tokens can be words, sentences, or paragraphs. Tokenization can separate existing writing by looking for or finding delimiters in a word. Common or usual word boundaries are spaces or Whitespace.

examples :
Before Tekonize
“Adapun situasi setiap orang berbeda”
After Tokenize
“Adapun” ”situasi” ”setiap” “orang” “berbeda” = Number of Tokens 7

5) Feature Selection

Feature Selection is the selection of attributes that are necessary before the data analysis process is carried out. This process occurs because there is a lot of data available and the amount will continue to increase, so it is necessary to carry out Feature Selection in order to increase the value of accuracy, as well as improve performance on data with very high dimensions.

Apart from that, Feature Selection attributes are carried out so that the form of the data to be analyzed becomes simpler and also reduces the size of the data that will be processed during data collection. Examples can be seen in Fig. 2

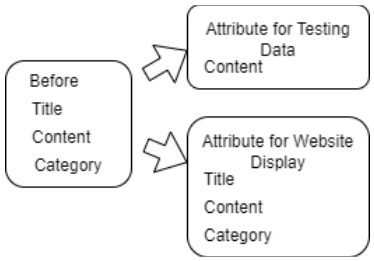

Fig. 2 Feature Selection

C. Entity Relationship Diagram (ERD)

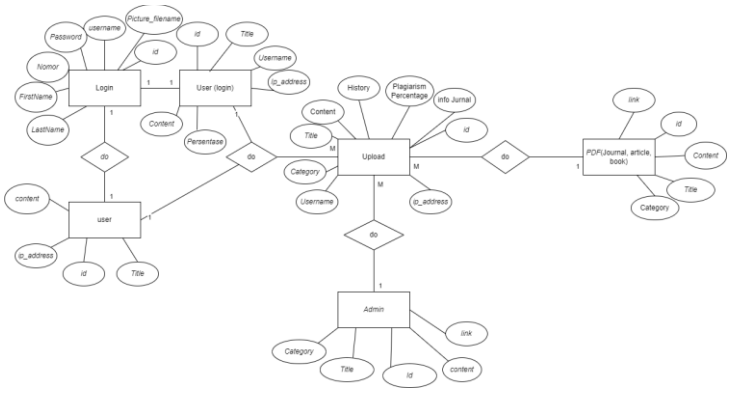

Fig. 3 Entity Relationship Diagram

An Entity Relationship Diagram (ERD) is a visual representation that identifies the types of entities in a system, describes their attributes, and illustrates the relationships between these entities [15]. An example of an ERD can be seen in Fig. 3

D. Data Flow Diagram (DFD)

DFD is a data flow-oriented system design tool which has a decomposition concept that can be used to explain aspects of system design and analysis that are easy to communicate to system users[16]. In designing a DFD there are several guidelines on how to describe a DFD well, namely:

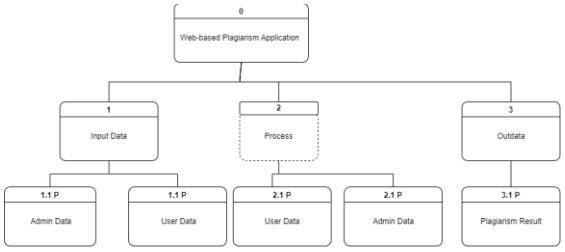

Fig. 4 Hierarchy Chart

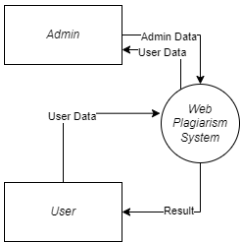

Fig. 5 Context Diagram

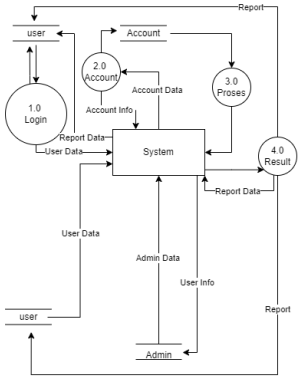

Fig. 6 Overview Diagram

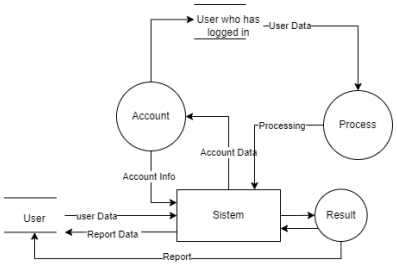

Fig. 7 Level 1

1) *Hierarchy Chart*
   Hierarchy Chart are usually used to prepare depictions from DFD to lower levels. Multilevel charts can be depicted using the process notation commonly used in DFD[17]. Can be seen in Fig. 4
2) *Context Diagram (Top Level)*
   Context Diagram is a diagram that consists of processes and represents the scope of the system. Context diagram is the highest level of DFD. Context diagram usually describe system input or output in a system [16]. Can be seen in Fig. 5
3) *Overview diagram (Level 0)*
   Overview Diagram is a picture that provides a brief and comprehensive view of the system covered, a picture that shows the main functions or processes involved in the system [16]. Can be seen in Fig. 6
4) *Detailed Diagram (Level 1)*
   Detailed Diagram is a process that is broken down from level 0 into more detailed processes [17]. Can be seen in Fig. 7

E. Blackbox Testing

Black box testing is a test that is created as a result of the execution of an application through tested data to ensure the functionality of the application being run. This process is carried out to verify whether the application is in accordance with the requirements or not [18]. In this black box there are several testing methods, such as Sample Testing, Boundary Value Analysis, and Equivalence Partitions.

In this research, the black box testing method that will be used is Equivalence partitions. Equivalence partitions is a testing process based on data input in each form created in the application, each input menu will be tested and will be grouped according to its function, whether it is valid or invalid [19]. An example of the black box Equivalence partitions technique can be seen in TABEL 3.

TABLE 3
BLACKBOX TESTING

No.	Form/Tested Display	Number of Test Items	Result		Amount of Evidence
			True	False	
1	Login	2	2	0	2
2	Register	2	2	0	2
3	Menu user has not logged	5	5	0	5

4	Menu user has not logged	5	5	0	5
5	Menu User Account Setting	1	1	0	1
6	Menu Admin	10	10	0	10
7	Menu Datasets	10	10	0	10
8	Menu Edit Dataset	10	10	0	10
9	Menu User Login	2	2	0	2
10	Menu Edit User Login	2	2	0	2
11	Menu Admin History who has logged in	5	5	0	5
12	Admin Menu Edit Login History	5	5	0	5
13	Admin History menu for those who have not logged in	5	5	0	5
14	Admin Menu Edit History for those who are not logged in	5	5	0	5
15	Results Menu	5	5	0	5

F. TF-IDF

TF-IDF (term frequency–inverse document frequency) is a calculation of the frequency of appearance of a term in existing documents. The TF-IDF algorithm will check the appearance of each word in the document from the results of filtering and tokenization. The TF-IDF formula can be calculated as follow :

$$W_{ij} = tf_{ij} * idf \quad (1)$$

Note :
W_{ij} : Word weight of the j-term and i-document
tf_{ij} : The number of occurrences of the j word/term in the i document
idf : log N/df
N : The number of all existing documents
n : Number of documents containing the j term

Example can be seen in TABLE 4 and TABLE 5:
D1:Aku Bisa
U :Bisa Kenapa

TABLE 4
PROCESS TO FIND IDF

Tokens	u	D1	df	IDF LOG (N/df)
Aku	0	1	1	$\frac{2}{1}$ = 0.301029996
Bisa	1	1	2	$\frac{2}{2}$ = 0
Kenapa	1	0	1	$\frac{2}{1}$ = 0.301029996

TABLE 5
TF-IDF

Tokens	u	D1
Aku	0	0.301029996
Bisa	0	0
Kenapa	0.301029996	0

G. Cosine Similarity

Cosine Similarity in this research is used for the process of calculating the similarity of documents. The formula for Cosine Similarity is as follow:

$$SV = \frac{AB}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

Note :
A : Vector A
B : Vector B
||A|| : Vector A length
||B|| : Vector B length

Example :

$$D1 = \frac{(0*0.301029996)+(0*0)+(0.301029996*0)}{\sqrt{0^2+0^2+0.301029996^2}*\sqrt{0.301029996^2+0^2+0^2}} = 0$$

H. Text Mining

Text Mining is the process of examining large amounts of documents to find new information or help answer specific research questions. identify the remaining facts, relationships, and statements. Once extracted, this information is converted into a structured form that can be further analyzed or presented directly using grouped HTML tables, mind maps, charts, etc. [8]

In this research, text mining is at the core of everything because with the text mining process all data can be obtained and preprocessed, even though text mining only has 2 words, the world of text mining is very wide. There are many uses that can be obtained from text mining. Text mining consists of 3 main tools [20], namely:

1) *Extraction of Information*
   This step is carried out by following a predetermined text structure by matching the pattern. This step is the initial step where a writer extracts relevant information from an unstructured research article. Writers usually select appropriate articles, and try to find and gather information by distinguishing relevant texts.
2) *Discovery of Suitable Text*
   Discovery of Suite Text (DoscoText) is the most important aspect of text mining. Why because DoscoText is a collection of structured data from unstructured text. Usually authors use knowledge discovery in database (KDD) tools or knowledge discovery from databases to obtain structured and relevant data that can help in subsequent analysis. Writers use keyword extraction to classify text, create groups of terms, and so on.
3) *Text Analytics*
   Text Analytics is an automated process that helps interpret large amounts of unstructured text into qualitative data, primarily to uncover insights, trends, and patterns.

IV. RESULTS

A. Main Web Appearances

1) *Users who have logged*

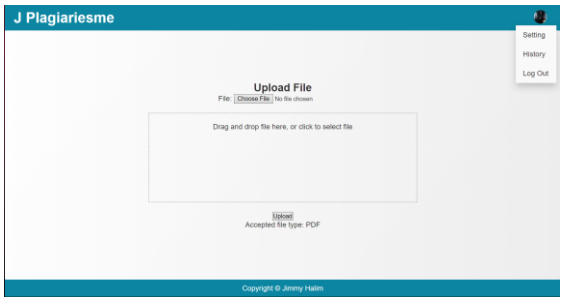

Fig. 8 Users who have logged

Users who have logged can be seen in Fig. 8, the menu display design of Users who have logged will look similar to the User has not logged display, the difference is the settings menu which will be visible on the right, the settings menu will open when the user's mouse points to the circle on the right which will be the place for profile photo later. The menu design for users who have logged in will consist of 2 menus which will take the user to other menus that have been provided.

2) *User has not logged*

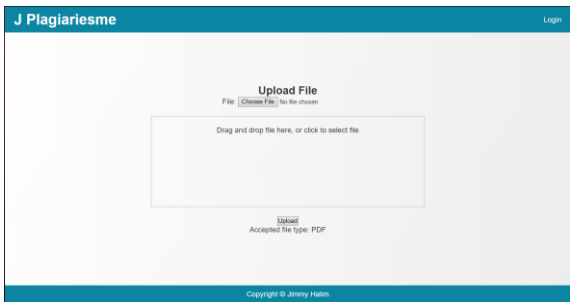

Fig. 9 User has not logged

The User has not logged menu will look like Fig. 9, where there will be a place to store files and an upload button which, when clicked, will carry out the process of retrieving data from the journal which will be stored in the database and will be used again for the next process.

B. Manual Calculation Results

Manual calculations will start from the TF-IDF process, where the results of the TF-IDF will be used again to find Cosine Similarity

1) TF-IDF

TABLE 6
IDF TABLE

Tokens	u	d1	d2	d3	df	idf
era	1	0	0	0	1	0.60206
globalisasi	1	0	0	0	1	0.60206
ini	1	0	0	0	1	0.60206
kembang	1	0	1	0	2	0.30103
pola	1	0	0	0	1	0.60206
hidup	1	0	0	0	1	0.60206
masyarakat	1	0	0	0	1	0.60206
plagiarisme	0	1	0	0	1	0.60206
tindak	0	1	0	0	1	0.60206
langgar	0	1	0	0	1	0.60206
momok	0	1	0	0	1	0.60206
ilmu	0	1	1	0	2	0.30103
tahu	0	1	0	0	1	0.60206
laku	0	1	0	0	1	0.60206
lapor	0	0	1	0	1	0.60206
teliti	0	0	1	0	1	0.60206
fakultas	0	0	1	0	1	0.60206
dana	0	0	1	0	1	0.60206
boptn	0	0	1	0	1	0.60206
tahun	0	0	1	0	1	0.60206
anggar	0	0	1	0	1	0.60206
aplikasi	0	0	0	1	1	0.60206
terap	0	0	0	1	1	0.60206
algoritma	0	0	0	1	1	0.60206
cocok	0	0	0	1	1	0.60206
string	0	0	0	1	1	0.60206
dokumen	0	0	0	1	1	0.60206

TABLE 7
RESULTS TF-IDF TABLE

Tokens	u	d1	d2	d3
era	0.60206	0	0	0
globalisasi	0.60206	0	0	0
ini	0.60206	0	0	0
kembang	0.30103	0	0.30103	0
pola	0.60206	0	0	0
hidup	0.60206	0	0	0
masyarakat	0.60206	0	0	0
plagiarisme	0	0.60206	0	0
tindak	0	0.60206	0	0
langgar	0	0.60206	0	0
momok	0	0.60206	0	0
ilmu	0	0.30103	0.30103	0

tahu	0	0.60206	0	0
laku	0	0.60206	0	0
lapor	0	0	0.60206	0
teliti	0	0	0.60206	0
fakultas	0	0	0.60206	0
dana	0	0	0.60206	0
boptn	0	0	0.60206	0
tahun	0	0	0.60206	0
anggar	0	0	0.60206	0
aplikasi	0	0	0	0.60206
terap	0	0	0	0.60206
algoritma	0	0	0	0.60206
cocok	0	0	0	0.60206
string	0	0	0	0.60206
dokumen	0	0	0	0.60206

The TF-IDF process can be searched by looking for the IDF value first, can be seen in TABLE 6. , after the IDF is found the next process is multiplying by tf as shown in TABLE 7. This process is carried out to obtain the results of the word weighting process from the words in the user documents provided. After the TF-IDF process has been carried out, the next process is to calculate the Cosine Similarity from the TF-IDF results.

2) Cosine Similarity

The Cosine Similarity calculation process can be carried out with the formula:

$$SV = \frac{AB}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

TABLE 8
MANUAL CALCULATION RESULTS

Documents	Result
D1	0
D2	4%
D3	0

C. Web Result View

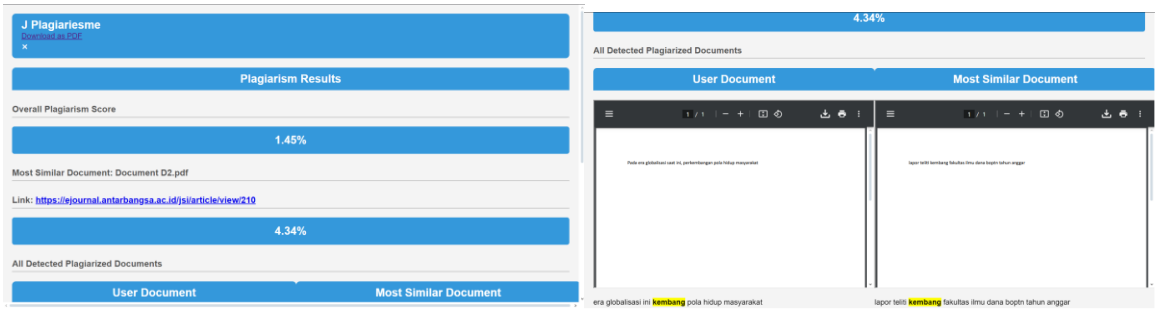

Fig. 10 Web Application Result View

The results menu can be seen in Fig. 10 here is a display of the results of the calculation processes that have occurred in the system, here the plagiarism results will be displayed in percentage form, where the display will show the plagiarism results from the sum of all PDF data in the database and divided with the amount of PDF data in the database, plagiarism result is the highest among the data in the database, and will display any words that are plagiarized in color.

V. DISCUSSION

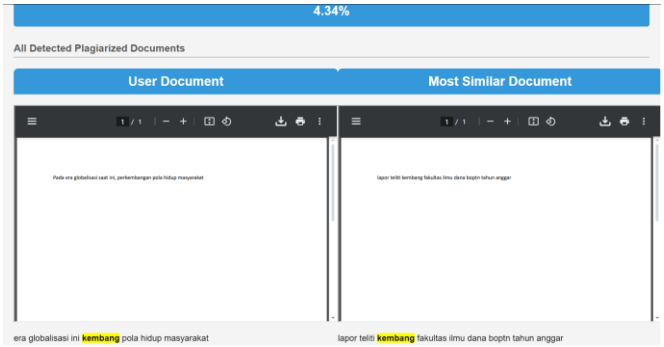

Fig. 11 Web Application Result

TABLE 9
Manual Calculation Results

Documents	Result
D1	0
D2	4%
D3	0

The methods used in this journal are TF-IDF and Cosine Similarity, from the results of manual calculations and web applications, the process results can be said to be almost similar to the application value of 4.34% which can be seen in Figure 11. and the manual calculation of 4% can be seen in TABLE 9, these two results can be said to be accurate because the results are not too far apart.

Using the TF-IDF and Cosine Similarity techniques has its own advantages and disadvantages. The advantages of TF-IDF and Cosine Similarity can be said to be more accurate because it uses 2 techniques with TF-IDF which calculates word weights from the contents of each document and Cosine Similarity which calculates the similarity of each document from the TF-IDF results, however, there are disadvantages of the TF-IDF process IDF and Cosine Similarity will take a lot of time in the program process and manual calculation process if you have a lot of data.

The plagiarism process uses Indonesian. By using Indonesian, this application can run well and as expected. but what about the concept of multilingualism, no further research has been carried out on the concept of multilingualism. With the multilingual concept, this application will be able to run efficiently, but new concepts are also needed to adopt multilingualism, such as running coding. However, the concept of multilingualism can be used as a good opportunity for future development.

VI. CONCLUSIONS

Based on the experiments carried out, it can be concluded that this application which uses TF-IDF and Cosine Similarity works effectively and runs smoothly. with the experiments carried out it can be said that this application was successfully created, but even so there is something that needs to be improved again, namely the process of the TF-IDF and Cosine Similarity programs which can be said to take quite a long time in processing calculations.

Suggestions that can be given for further research are to look for special programs that can shorten the program runtime or compare the process runtime of this method with other calculation methods. You can also use NLP-based approaches such as Word2Vec or BERT for more efficient and sophisticated detection. and you can also carry out further research for multi-format file processes, such as doc and others.

REFERENCES

[1] L. Hermawan and M. B. Ismiati, "Aplikasi Pengecekan Dokumen Digital Tugas Mahasiswa Berbasis Website," *J. Buana Inform.*, vol. 11, no. 2, pp. 94–103, 2020, doi: https://doi.org/10.24002/jbi.v11i2.3706.

[2] M. A. Shadiqi, "Memahami dan Mencegah Perilaku Plagiarisme dalam Menulis Karya Ilmiah," *Bul. Psikol.*, vol. 27, no. 1, p. 30, Jun. 2019, doi: 10.22146/buletinpsikologi.43058.

[3] M. Krokoscz, "Plagiarism in articles published in journals indexed in the Scientific Periodicals Electronic Library (SPELL): a comparative analysis between 2013 and 2018," *Int. J. Educ. Integr.*, vol. 17, no. 1, p. 1, Dec. 2021, doi: https://doi.org/10.1007/s40979-020-00063-5.

[4] M. A. Pratiwi and N. Aisya, "Fenomena plagiarisme akademik di era digital," *Publ. Lett.*, vol. 1, no. 2, pp. 16–33, Jul. 2021, doi: 10.48078/publetters.v1i2.23.

[5] A. Riyani, M. Z. Naf'an, and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *J. Linguist. Komputasional*, vol. 2, no. 1, pp. 23–27, 2019, Accessed: May 07, 2024. [Online]. Available: https://scholar.archive.org/work/7t7hzdt6gnbg5e7npnbdsuoobi/access/wayback/http://inacl.id:80/journal/index.php/jlk/article/download/17/19/

[6] Herianto, Yulisman, W. H. Manullang, and Y. Irawan, "APLIKASI DETEKSI PLAGIARISME JUDUL TUGAS AKHIR BERBASIS WEB DENGAN MENGGUNAKAN ALGORITMA RABIN-KARP ROLLING HASH (STUDI KASUS: AMIK MAHAPUTRA RIAU)," *J. ILMU Komput.*, vol. 10, no. 2, pp.

107–112, 2021, doi: https://doi.org/10.33060/JIK/2021/Vol10.Iss2.223.

[7] K. Abdullah *et al.*, *Metodologi Penelitian Kuantitatif*, vol. 3, no. 2. Pidie Provinsi Aceh: Yayasan Penerbit Muhammad Zaini, 2022. [Online]. Available: https://www.infodesign.org.br/infodesign/article/view/355%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/731%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/269%0Ahttp://www.abergo.org.br/revista/index.php/ae/article/view/106

[8] I. Tojimamatov, T. D. Abdusalomovna, H. O. A. Qizi, K. O. A. Qizi, and M. D. B. Qizi, "TEXT MINING," *Eur. J. Interdiscip. Res. Dev.*, vol. 13, pp. 184–189, 2023.

[9] D. Remawati and H. Wijayanto, *Buku Ajar Web Jsp Dengan Database Mysql*. SEMARANG: LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT UNIVERSITAS DIAN NUSWANTORO SEMARANG, 2021. [Online]. Available: https://eprints.sinus.ac.id/784/1/Buku_Ajar_Web_JSP_dengan_database_MySQL.pdf

[10] R. A. Putri, *Buku Ajar BASIS DATA*, 2nd ed. Medan: PENERBIT MEDIA SAINS INDONESIA (CV. MEDIA SAINS INDONESIA), 2022.

[11] M. Madani and Haryono, "Perancangan Website Media Promosi Produk Gerabah Menggunakan Metode Waterfall Designing a Promotional Media Website for Pottery Products Using the Waterfall Method," *J. Bumigora Inf. Technol.*, vol. 5, no. 2, pp. 195–204, 2023, doi: https://doi.org/10.30812/bite/v5i2.3370.

[12] C. Ningki and Noviyanti, "Implementasi Aplikasi Penjualan Produk Tradisional Berbasis Website Menggunakan Metode Waterfall," *J. Inform.*, vol. 19, no. 2, pp. 107–114, 2023, doi: https://doi.org/10.52958/iftk.v19i2.6149.

[13] C. Zong, R. Xia, and J. Zhang, *Text Data Mining*. China: the registered company Springer Nature Singapore, 2021. doi: https://doi.org/10.1007/978-981-16-0100-2.

[14] A. Hermawan, I. Jowensen, Junaedi, and Edy, "Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 1, pp. 129–137, Apr. 2023, doi: https://doi.org/10.23887/jstundiksha.v12i1.52358.

[15] F. N. Hasanah and R. S. Untari, *REKAYASA PERANGKAT LUNAK*. Sidoarjo: UMSIDA Press, 2020.

[16] F. Wahyuni, "PERANCANGAN SISTEM INFORMASI KAS BERBASIS WEB DENGAN MENGGUNAKAN METODE WATERFALL," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 7, no. 1, pp. 138–143, Apr. 2023, doi: https://doi.org/10.46880/jmika.Vol7No1.pp138-143.

[17] A. Rozaq, *KONSEP PERANCANGAN SISTEM INFORMASI BISNIS DIGITAL*. Banjarmasin,: Poliban Press, 2020. [Online]. Available: https://repository.stkipjb.ac.id/index.php/lecturer/article/view/3694/3111

[18] M. N. Huda, M. Burhan, A. Satibi, H. A. Pradita, A. Saifudin, and I. Kusyadi, "Implementasi Black Box Testing pada Aplikasi Sistem Kasir dengan Menggunakan Teknik Equivalence Partitions," *J. Teknol. Sist. Inf. dan Apl.*, vol. 5, no. 2, pp. 120–127, 2022, doi: https://doi.org/10.32493/jtsi.v5i2.17645.

[19] M. Y. Suyudi, A. P. Pratiwi, R. F. Mawahdah, Y. A. Purwara, and I. Kusyadi, "Teknik Pengujian Equivalents Partitioning pada Aplikasi Sistem Pendaftaran PAUD berbasis WEB dengan Menggunakan Black Box," *J. Inform. Univ. Pamulang*, vol. 5, no. 2, pp. 198–202, 2020, doi: https://doi.org/10.32493/informatika.v5i2.5351.

[20] V. Puri, S. Mondal, S. Das, and V. G. Vrana, "Blockchain Propels Tourism Industry—An Attempt to Explore Topics and Information in Smart Tourism Management through Text Mining and Machine Learning," *Informatics*, vol. 10, no. 1, p. 9, Jan. 2023, doi: 10.3390/informatics10010009.