Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 7 No. 1, April 2022, pp. 44 – 57

Cross-Language Text Document Plagiarism Detection System Using Winnowing Method

Mustika Mentari *[1], Imam Fahrur Rozi [2]
State Polytechnic of Malang, Jl. Soekarno Hatta No.9, Jatimulyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65141, (0341) 404424-404425 / (0341) 404424
*E-mail : must.mentari@polinema.ac.id *[1], imam.rozi@polinema.ac.id [2]*
**Corresponding author*

Maria Puji Rahayu
State Polytechnic of Malang, Jl. Soekarno Hatta No.9, Jatimulyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65141, (0341) 404424-404425 / (0341) 404424
E-mail : mariapujir@gmail.com [3]

Abstract - Currently, there are many text documents such as journals scattered on the internet, both Indonesian and English-language journals. With this, it is possible to act plagiarism by copying from foreign journals that are translated into other languages or copying directly without being changed from the original language. One way that can suppress these actions is to build a plagiarism detection system for cross-language text documents. The method that can be used to detect document plagiarism is the Winnowing method. Winnowing method is a method where text input will be processed to produce a hash value called a fingerprint. This study aims to build a system that can detect plagiarism of text documents in different languages using the Winnowing method. Text documents that can be tested are input text and PDF files. Documents used in system testing are journals that have the same topic. The results of the highest level of accuracy produced between the calculation of the Jaccard Coefficient with the Plagiarism Checker X application are in the fourth scenario with an average percentage value of 84.7%.

Keywords - plagiarism, winnowing method, jaccard coefficient

1. INTRODUCTION

Cross-language plagiarism is copying someone else's work from a different language and translating it into another language [1]. Detecting plagiarism across languages requires additional components to translate each sentence. The translation results must be accurate, in order to optimizing the final result of detection of similarity between documents. Such as research conducted by Putri Ratna who uses the Googletrans API to translate English text documents into Indonesian [2].

When detecting plagiarism of scientific works it is usually done manually namely by matching new scientific works with existing scientific works, cara seperti ini dianggap tidak efisien [3]. Therefore, a cross-language text document plagiarism detection system is needed. One method that can be used to detect the similarity of text documents is the Winnowing method [4].

Winnowing method is a method where text input will be processed so as to produce a collection of hash values called fingerprint [5]. The hash value is a numeric value obtained from

ASCII calculations [6]. The accuracy results of the Winnowing method are better than the rabin carp method, which produces an accuracy rate of 88.89% and a processing time of 0.13 seconds, while the rabin carp method produces an accuracy rate of 37.50% with a processing time of 0.19 seconds [7]. Several previous research groups discussing the same subject with the title can be seen in Table 1

Table 1. Previous Research

No	Source	Problem	Method	Result
1	[8]	The occurrence of acts of plagiarism in writing essays	TF-IDF	Successfully detected the similarity of the tested document and the comparison document which resulted in a 0% difference between the system and manual calculations
2	[9]	There is no automatic plagiarism detection yet that can check quotes from foreign languages.	Winnowing	The resulting accuracy results are higher if you use an English-Indonesian dictionary than if you don't use an English-Indonesian dictionary
3	[7]	There are acts of plagiarism against other people's work freely via the internet	Rabin Karp and Winnowing	The Rabin Karp algorithm has an accuracy of 37.5% with a processing time of 0.19 seconds while the Winnowing algorithm produces an accuracy of 88.89% with a processing time of 0.13 seconds.

From the explanation above, we propose a similarity detection system between cross-language text documents using the Winnowing method because the Winnowing method has a better accuracy rate than the Rabin Karp method. The languages that can be detected in text documents are English and Indonesian. The English text document will be translated into Indonesian and then preprocessed will be carried out, forming the n-gram value, calculating the hash value, forming a hash window, forming a fingerprint and calculating the percentage similarity value using the Jaccard Coefficient. Therefore, a study was made entitled "Plagiarism Detection System of Cross-Language Text Documents Using the Winnowing Method".

2. RESEARCH METHOD

2.1. Plagiarism

Plagiarism is a dishonest act caused by a lack of creativity and innovation in creating original or original works [10]. In looking for similarities between text documents such as journals, plagiarism checks are carried out which are useful for checking the authenticity of the journals [3]. There are three methods for detecting plagiarism, namely the full text comparison method, the document fingerprint method, and the keyword similarity method [11]. The law that regulates the Prevention and Handling of Plagiarism in Higher Education is regulated in Law no. 17 of 2010. In Chapter II Article 3 it is written that what includes plagiarism in higher education is one or more students, one or more lecturers/researchers/educational staff or one or more lecturers/researchers/educational staff together with one or more students. full text comparison, document fingerprint method and keyword similarity method [11].

2.2. Text Mining and Text Preprocessing

Text mining is the process of finding patterns or extracting information from text data to generate new information [12]. The data source used in text mining is a collection of text that has an unstructured or at least semi-structured format [13]. Text preprocessing is a

process that is often used in text mining [14]. There are 4 types of preprocessing stages including case folding, filtering, stemming, and tokenizing [15]. However, in this study only three preprocessing stages were carried out as shown in Figure 1 below.

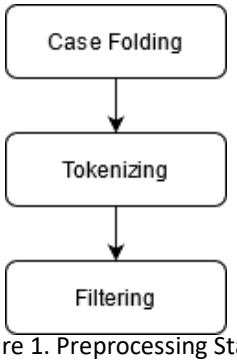

Figure 1. Preprocessing Stage

2.3. Winnowing method

Winnowing method is a method which takes the important attributes of each fingerprint to detect similarities between sentences [16]. Fingerprint is the result of a set of hash values that are used as a basis for comparison between text files [15]. This method was first introduced in 2003 by Schleimer et al. with a journal entitled Winnowing: Local Algorithms for Document Fingerprinting.

The steps in calculating the similarity of the text to Winnowing are as follows [17].

1) The first step is to remove unimportant characters such as punctuation marks, spaces, and numbers. However, in this study there are additional processes of case folding, tokenizing, and filtering or stopword removal. In stopword removal, it is useful to remove irrelevant words. For example, in the sentence “Hari ini cuacanya cerah " it will be changed to *cuacanyacerah*.
2) The second stage is to form a series of grams from the results of the previous process. N-gram is the result of the cut of each character. The difference for the results of each n-gram lies in the distribution of characters. If n is used 2 then the word will be divided by 2 characters, whereas if n is 3 then the word will be divided by 3 characters [18]. An example of the results of the formation of an n-gram circuit with a value of n = 3 is as follows: *cua uac aca can any nya yac ace cer era rah.*
3) The third step is to calculate the hash value of each gram using a rolling hash. Rolling hash is a way to transform a string into a unique value with a certain length and serves as a marker for the string. In calculating the hash value using the following equation (1).

$$H_{(ck)} = c1 * b^{(k-1)} + c2 * b^{(k-2)} + \cdots + ck * b^{(k-k)} \quad (1)$$

Where:
c = ASCII code of the character
b = base (prime number)
k = many characters

4) Then the next step is to form a window from the hash value. Window is a division or grouping of several hash values with a specified size.

5) The last step is to choose the smallest value from each window to be used as a fingerprint. Fingerprint is the result of forming a window from the selection of the smallest hash value [9].

2.5 Jaccard Coefficient

Jaccard Coefficient is an equation that is used to calculate the similarity (similarity) of the results of fingerprint formation, so the percentage of text similarity is produced [4]. The Jaccard Coefficient equation is written in equation (2) below:

$$Similarity\ precentage = \frac{The\ same\ number\ of\ fingerprints}{Total\ all\ fingerprints} \times 100\% \qquad (2)$$

2.6 Dataset

In the data collection method, literature studies and observations were carried out. In the literature study, an understanding of the Winnowing method was carried out through articles, literature, and books. After that, observations were made by looking for and observing existing journals on the sites ieexplore.ieee.org, njca.co.id, link.springer.com, alweb.org, citeseerx.ist.psu.edu, stmikplk.ac.id, cursorjournal.org, researchgate.net and jip.polinema.ac.id. The journal will be used as test data.

2.7 System Design

As seen in Figure 2, the first process to detect plagiarism is to enter the document to be tested in the form of a text or PDF file and will be stored in the database. In addition, users can also compare test documents with more than one comparison document that is already stored in the database. The inputted text will be detected in the language used. If the text is in English, it will be translated first to Indonesian. The contents of the translated document will be given a test limit by not including sentences in quotation marks, citations, and references. After that, it will be continued with the process of changing the sentence into lowercase along with the removal of characters and words that are not important. The process carried out is preprocessing. The results of the preprocessing will be processed to produce fingerprints using the Winnowing method. Winowwing starts from the formation of the n-gram value and continues by calculating the hash value of each n-gram. Then a window is formed from the hash value and after that proceed with the selection of fingerprints from each window. The fingerprint results obtained will be searched for the percentage of similarity using the jaccard coefficient. The percentage value of the jaccard coefficient is the final value of the similarity between documents.

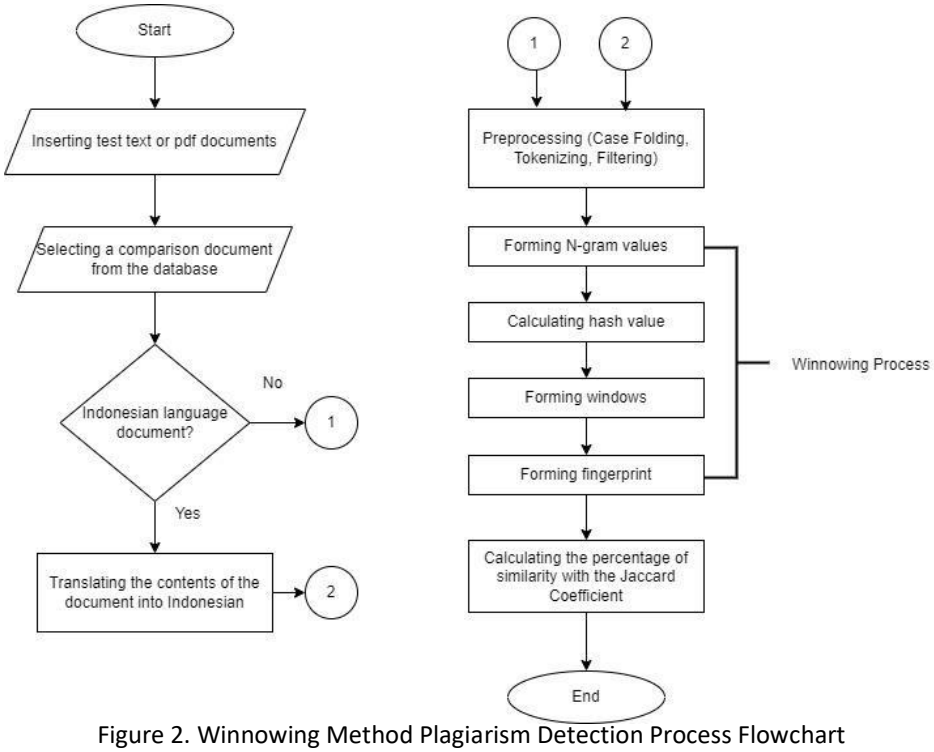

Figure 2. Winnowing Method Plagiarism Detection Process Flowchart

2.8 System Testing Method

In the test, there are 3 stages of testing, namely functional testing, system validity testing and user testing. Testing the validity of the system is a test by comparing the percentage results from the calculation of the jaccard coefficient with the Plagiarism Checker X application which is carried out with 5 scenarios. In the first scenario the test and comparison documents are used in English, the second scenario uses Indonesian test documents and English comparison documents, the third scenario uses Indonesian English test and comparison documents, the fourth scenario uses Indonesian test and comparison documents, and the fifth scenario uses Indonesian language test document and Indonesian English language comparison document.

3. RESULTS AND DISCUSSION

3.1 Database Implementation

In Figure 3 is a database design that was built where there is a relationship between the tb_hitung table with tb_document and tb_document with the user. Fill in the id_test and id_banding columns in the tb_hitung table based on the contents of the id_document column in the tb_document table. The id_user column in the tb_document table is based on the contents of id_user in the user table. While the admin table does not have any relationship with other tables.The tb_hitung table is useful for storing test document data entered by users and admins. The tb_hitung table is useful for storing document calculation results. The user table is useful for storing user data, while the admin table is useful for storing admin data.

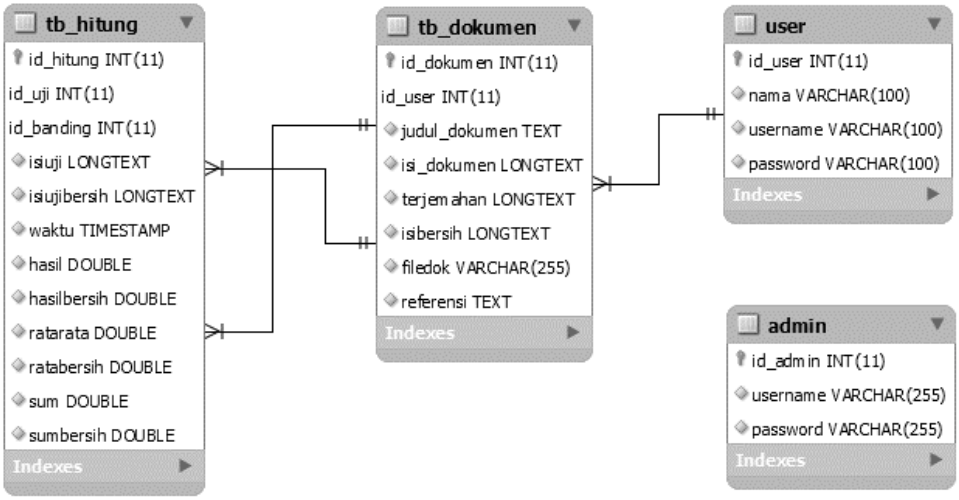

Figure 3. Database Design

3.2 System Implementation

In the implementation of this cross-language text document plagiarism detection system using the PHP programming language version 7.2.11.

1) Homepage

There are 2 home pages in the application created, namely the home page displayed on the user and the home page displayed on the admin. To be able to start doing calculations, users and admins must first login. The following is a display of the home page for user logins in Figure 4.

2) Document List Page

The personal document list page displays documents that have been successfully inputted by logged in users, both in the form of text and PDF. In the document content column, the contents of the entire document have not been processed at all, while in the translation column a translation has been carried out for the English text into Indonesian and has been given sentence boundaries before the abstract and bibliography have been deleted. The calculation action button is useful for selecting a document as a test document and will be redirected to the select comparison document page. The delete action button is useful for deleting the clicked document. The following is a view of the document list page for each user in Figure 5.

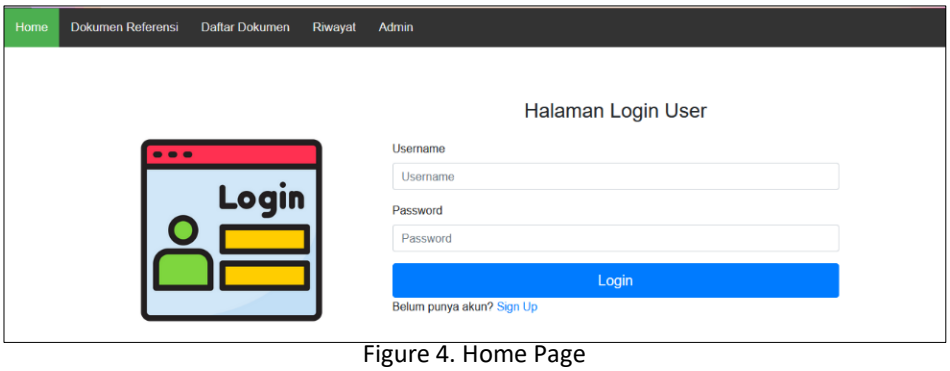

Figure 4. Home Page

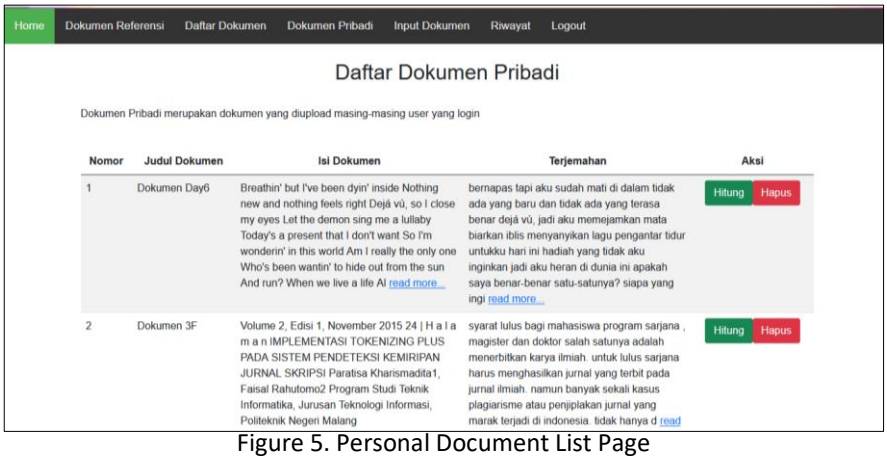

Figure 5. Personal Document List Page

In Figure 6 below is a page that displays a list of reference documents. This reference document is a document that has been calculated previously and produces a percentage of less than 15%. Reference documents can be used as comparison documents when detecting similarities between text documents.

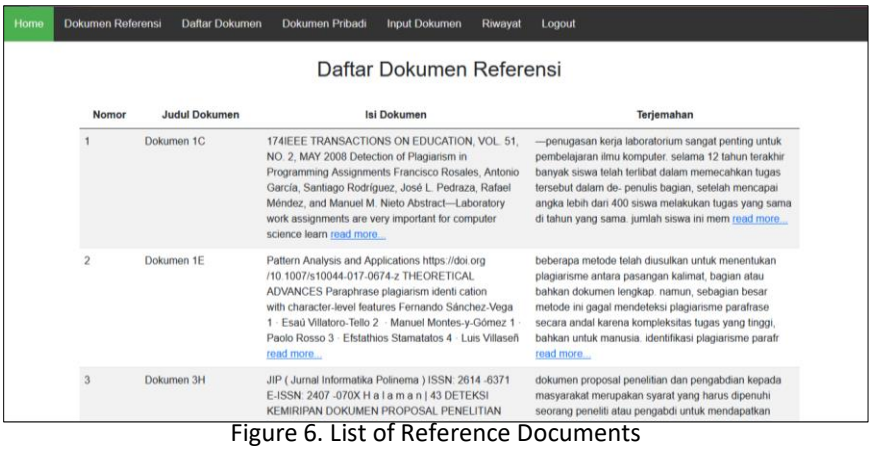

Figure 6. List of Reference Documents

3) Select Comparison Document Page

This comparison document select page can only be displayed if you have selected a test document when pressing the calculate button on the document list page. On this page it is useful to select the document to be compared. Users can choose a comparison document from personal documents or reference documents. If you have finished selecting the comparison document, you can press the calculate button and the page will switch to the calculation history. While the reset button is useful for deselecting the checkbox when selecting a comparison document. The following is a page display for selecting a comparison document from a reference document as shown in Figure 7.

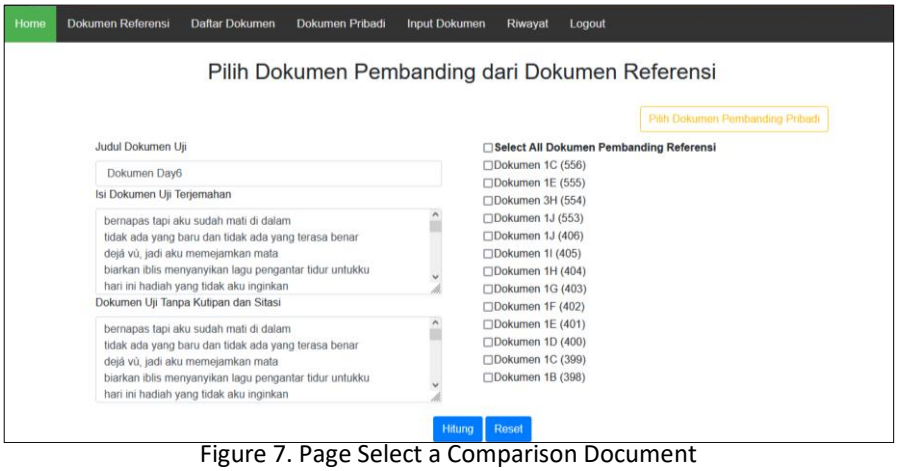

Figure 7. Page Select a Comparison Document

4) Document Input Page

On this page it is useful to insert a pdf document. The user must fill in the title of the document and the document file, so it can be saved. If one of the fields is not filled, it cannot be processed to be stored in the database. If everything is filled in, you can press the save button. If successful, the display will switch to the document list page. In Figure 8, the following is a display of the document input page in the form of a pdf to the user.

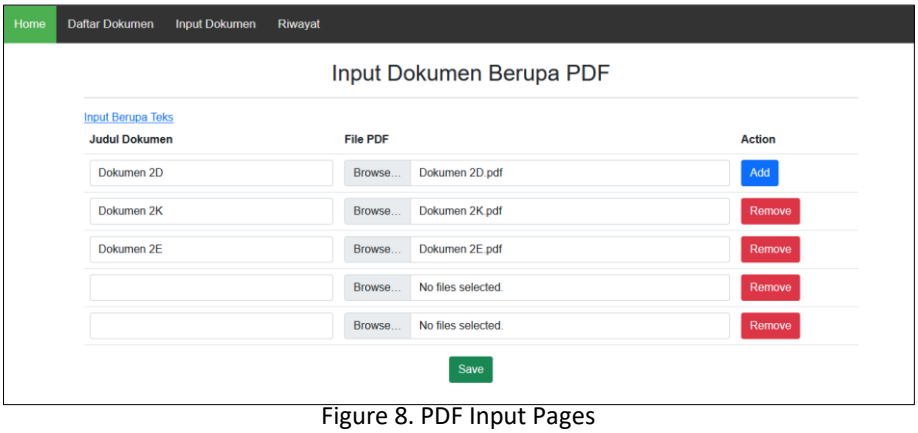

Figure 8. PDF Input Pages

5) Calculation History Page

On the calculation history page contains the time, test document id, number of appeal documents, results, and net results. The result is the sum of the equations between documents, while the net result is the sum of the results between documents but without quotes and citations. The delete button is useful for clearing the calculation history. Details of the equation calculation can be seen when pressing the calculate button and calculate net. The following is a display of the calculation history page for the user which can be seen in Figure 9.

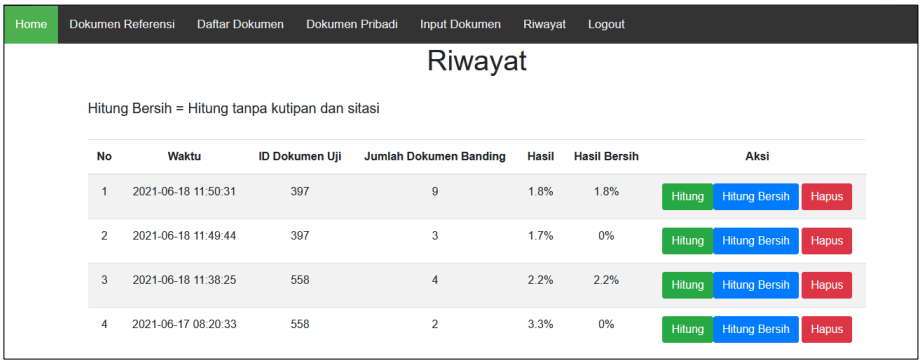

Figure 9. Calculation History Page

6) Calculation Result Detail Page

On the detail page, the calculation results display the result of the percentage of similarity using the Jaccard Coefficient of each document along with the sum of the similarities. The average result is the sum divided by the number of documents. However, the results displayed on the calculation history page are the results of the sum of the similarities of each document. The back button is useful for returning to the calculation history page. The following is a display of the detailed calculation results page for the user which can be seen in Figure 10.

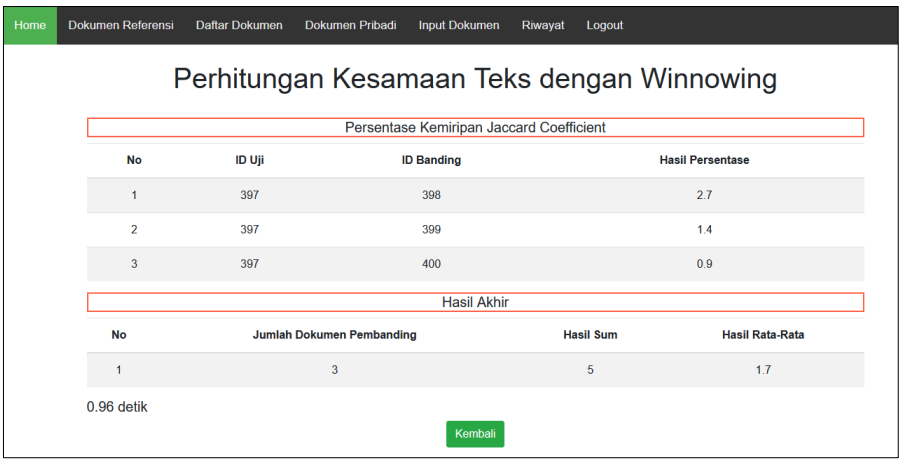

Figure 10. Calculation Result Detail Page

3.3 Test Result

Functional testing conducted with 22 test items using the black box method resulted that all test items were accepted or could be successfully carried out as planned. Validity of the plagiarism checker system using 32 Indonesian and English text documents with 5 scenarios.

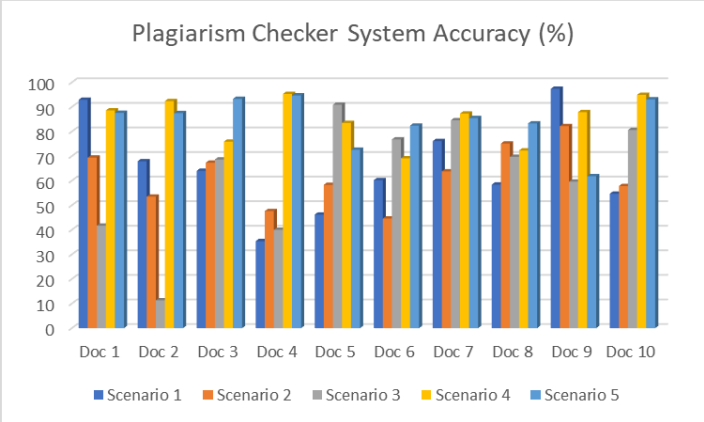

Figure 11. System Accuracy

The percentage of overall system success in the 5 scenarios that have been carried out is shown in figure 11. This percentage is obtained from the calculation of the difference between the results of checking plagiarism using the system that has been created (Jaccard Coefficient) and plagiarism checker X. For example, if the results of plagiarism checker X are 17 %, while the results of the Jaccard Coefficient are 15.8%. The number of appeal documents is 9 documents. Then the percentage of system success is calculated in the following way:

$$System\ Accuracy = \left(\frac{15.8}{9} / \frac{17}{9}\right) \times 100 = \left(\frac{1.75555}{1.88889}\right) = 92.9\%$$

The calculation of the difference in accuracy results or called the system accuracy system in the calculation will be used continuously in the explanation of figures 12 to figure 16. The numbers that appear in the series of chart figures 12 to figure 16 are rounded off from the accuracy results divided by the number of comparison documents. Supposed in this example 1.75555 can be rounded to 1.8.

In the first scenario, English document is compared to 9 English documents. The first test document and comparison document until the tenth document in scenario 1 are named 1A-1J. An example of the first stage is a comparison between document 1A and document 1B-1J, the second stage is a comparison between document 1B and 1A, 1C-1J, and so on. The document that produces the lowest level of system accuracy is document 4 with an average value of Jaccard Coefficient 9 documents of 1.8 (this number is the result of rounding) and the average result of plagiarism checker X is 1.9 (this number is the result of rounding). So, the percentage of success from checking plagiarism can be calculated from the difference in plagiarism detection using the two methods of 35.4% (system accuracy). While the highest level of system accuracy is document 10 with a value of 97.4%. The average level of system accuracy in this first scenario is 65.3%. Figure 12 below is a comparison chart of the percentage of document similarity in the first scenario.

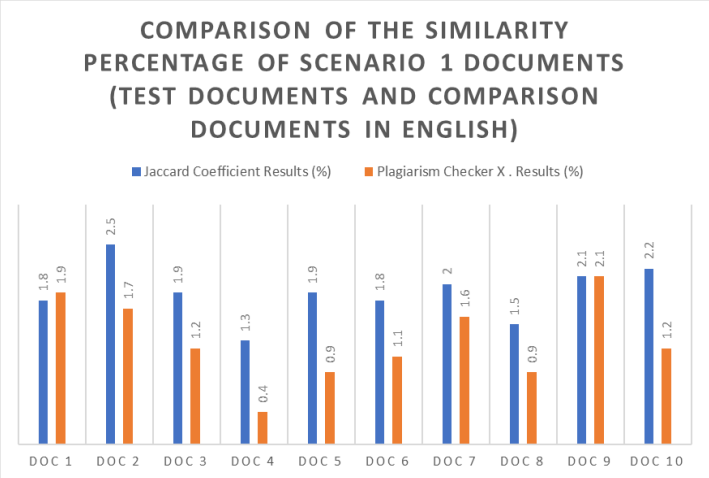

Figure 12. Graph of Document Similarity Percentage in Scenario 1

The test was carried out between 1 Indonesian document and 10 comparison documents in the second scenario. The test document in this scenario has a file name of 3A-3J, while the first comparison document to the tenth document in scenario 2 are named 1A-1J. An example of the first stage is a comparison between document 3A and document 1A-1J, the second stage is a comparison between document 3B and 1A-1J, and so on. The level of system accuracy that has the lowest results is in document 6 with a system success percentage of 44.6%. While the highest level of system accuracy is in the 9th Document with a percentage of 82.2%. The average level of system accuracy in this second scenario is 62%. In Figure 13, the following is a graph of the percentage of document similarity in the second scenario.

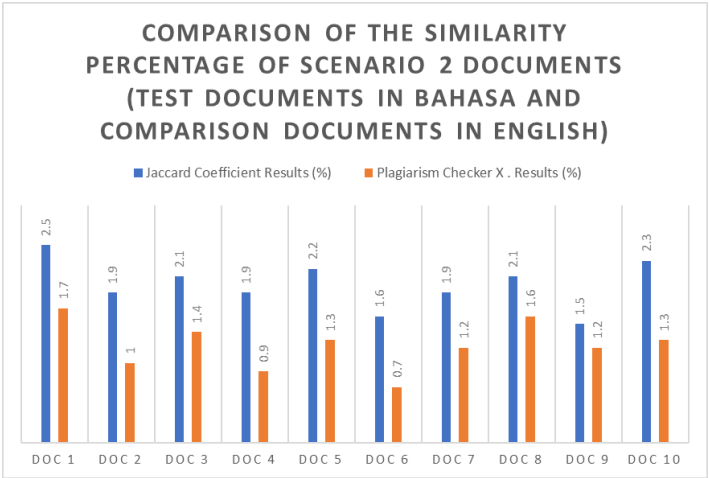

Figure 13. Graph of Document Similarity Percentage in Scenario 2

Figure 14 is a graph of the third scenario where the test is carried out between 1 Indonesian-English document and 9 Indonesian-English documents. The first test document and comparison document up to the tenth document in scenario 3 are named 2A-2J. An example of the first stage is a comparison between document 2A and document 2B-2J, the second stage is a comparison between document 2B and 2A, 2C-2J, and so on. The level of system accuracy that has the lowest value is in the second document with a system accuracy percentage of 11.3%. While the highest level of system accuracy is in the fifth document with a percentage of 90.9%. The results of the average level of system accuracy in this third scenario is 62.3%.

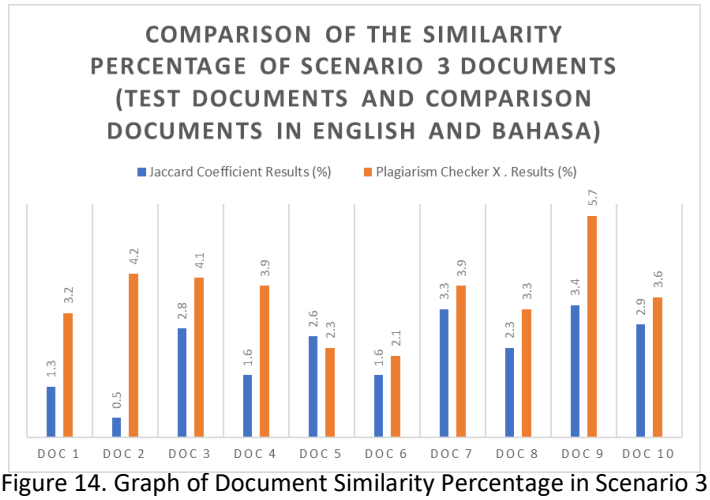

Figure 14. Graph of Document Similarity Percentage in Scenario 3

Figure 15 is a graph of the percentage of document similarity in the fourth scenario which compares 1 Indonesian document with 9 Indonesian documents. The first test document and comparison document up to the tenth document in scenario 4 are named document 3A-3J. Examples of the first stage are comparisons between documents 3A and documents 3B-3J, the second stage is a comparison between documents 3B and 3A, 3C-3J, and so on. The results of the lowest level of system accuracy produced are in the sixth document with a percentage of 69.1%. While the highest level of system accuracy is in the fourth document with a percentage of 95.3%. The average level of accuracy of the resulting system is 84.7%.

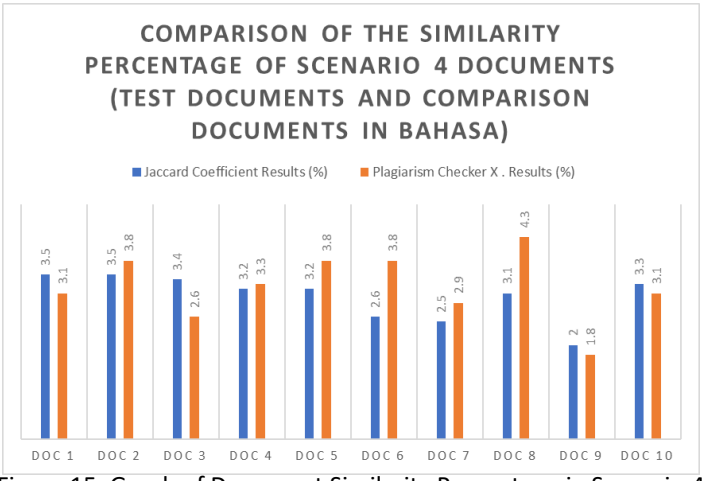

Figure 15. Graph of Document Similarity Percentage in Scenario 4

The fifth scenario compares 1 Indonesian document with 10 Indonesian documents in English. The test document in this scenario has a file name of 3A-3J, while the first comparison document to the tenth document in scenario 2 are named 2A-2J. An example of the first stage is a comparison between document 3A and document 2A-2J, the second stage is a comparison between document 3B and 2A-2J, and so on. The result of the lowest level of system accuracy in the comparison of the 9th document is 61.8%. While the highest level of system accuracy is produced in the fourth document with a percentage of 94.7%. The average level of system accuracy in this fifth scenario is 84.2%. In Figure 16, the following is a graph of the percentage similarity of the fifth scenario document.

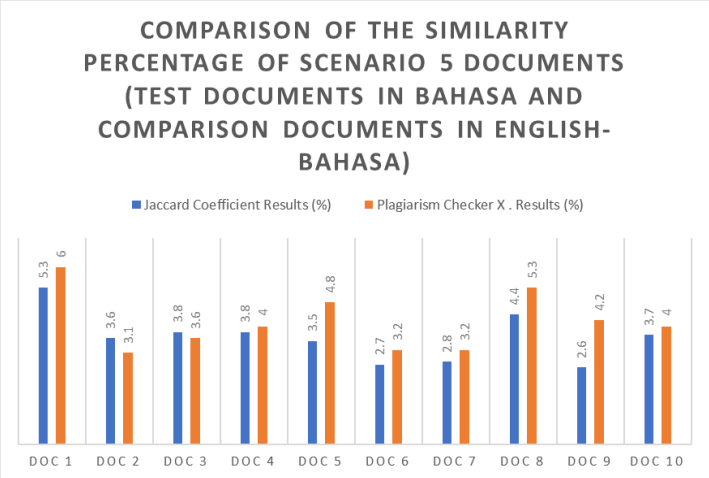

Figure 16. Graph of Document Similarity Percentage in Scenario 5

4. CONCLUSION

Based on the discussion and the results of the tests that have been carried out, it can be concluded that the plagiarism detection system of cross-language text documents that was built succeeded in processing texts in different languages, namely English and Indonesian. The success rate of the Winnowing method when applied to a cross-language text document plagiarism detection system that produces the highest average level of system accuracy is 84.7% by testing in scenario 4, while the second highest average level of system accuracy is 84.18% by testing in scenario 5. The results of this accuracy show that the system that has been created can detect plagiarism of documents effectively, both documents in the same language or across languages. In further research, it can be considered for document files that can be uploaded not only in PDF format but can be added in DOC or DOCX format.

REFERENCES

[1] B. Agarwal, "Cross-lingual plagiarism detection techniques for English-Hindi language pairs," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 22, no. 4, pp. 679–686, May 2019, doi: 10.1080/09720529.2019.1642626.

[2] A. A. Putri Ratna, F. Astha Ekadiyanto, I. Ibrahim, D. Husna, and F. Rahimullah, "Investigating Parallelization of Cross-language Plagiarism Detection System Using The Winnowing Algorithm in Cloud Based Implementation," in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, Oct. 2019, pp. 1–7. doi: 10.1109/ICAwST.2019.8923539.

[3] Pardede Jasman and Alvian Leo, "RANCANG BANGUN APLIKASI PENDETEKSI PLAGIARISME MENGGUNAKAN ALGORITMA SHERLOCK," *Jurnl Informatika*, vol. 6, no. 1, pp. 39–49, Jan. 2015.

[4] N. Alamsyah and M. Rasyidan, "DETEKSI PLAGIARISME TINGKAT KEMIRIPAN JUDUL SKRIPSI PADA FAKULTAS TEKNOLOGI INFORMASI MENGGUNAKAN ALGORITMA WINNOWING," *Technologia: Jurnal Ilmiah*, vol. 10, no. 4, p. 197, Oct. 2019, doi: 10.31602/tji.v10i4.2361.

[5] I. Ilham and P. Pasnur, "Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Pada Karya Tulis Mahasiswa," *Inspiration : Jurnal Teknologi Informasi dan Komunikasi*, vol. 7, no. 2, Dec. 2017, doi: 10.35585/inspir.v7i2.2447.

[6] I. W. S. Priantara, D. Purwitasari, and U. L. Yuhana, "Implementasi deteksi penjiplakan dengan algoritma winnowing pada dokumen terkelompok," Surabaya, 2011.

[7] 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.

[8] S. N. Lolyta, R. Y. Dillak, and F. E. Laumal, “Sistem deteksi plagiarisme lintas bahasa menggunakan algoritma tf-idf,” *Jurnal Ilmiah Flash*, vol. 5, no. 1, Jun. 2019.

[9] Sunardi, A. Yudhana, and I. A. Mukaromah, “PERANCANGAN APLIKASI DETEKSI PLAGIARISME KARYA ILMIAH MENGGUNAKAN ALGORITMA WINNOWING,” 2017.

[10] D. Leman, M. Rahman, F. Ikorasaki, B. S. Riza, and M. B. Akbbar, “Rabin Karp And Winnowing Algorithm For Statistics Of Text Document Plagiarism Detection,” in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Nov. 2019, pp. 1–5. doi: 10.1109/CITSM47753.2019.8965422.

[11] D. Susanto, A. Basuki, and P. Duanda, “Deteksi Plagiat Dokumen Tugas Daring Laporan Praktikum Mata Kuliah Desain Web Menggunakan Metode Naive Bayes,” *Nusantara Journal of Computersand its Applications*, vol. 2, no. 1, Dec. 2016.

[12] A. Indriani, A. Dahlan Jl Soepomo, and S. Janturan, *ANALISA KOREKSI KATA SOAL UJIAN SEMESTER DENGAN ALGORITMA LEVENSHTEIN DISTANCE*. 2018.

[13] T. Aprilianto and A. Badawi, “SISTEM KOREKSI KATA DAN PENGENALAN STRUKTUR KALIMAT BERBAHASA INDONESIA DENGAN PENDEKATAN KAMUS BERBASIS LEVENSHTEIN DISTANCE,” 2017.

[14] A. Filcha and M. Hayaty, “Implementasi Algoritma Rabin-Karp untuk Pendeteksi Plagiarisme pada Dokumen Tugas Mahasiswa,” *JUITA : Jurnal Informatika*, vol. 7, no. 1, p. 25, May 2019, doi: 10.30595/juita.v7i1.4063.

[15] R. Karisma Wibowo and K. Hastuti, “PENERAPAN ALGORITMA WINNOWING UNTUK MENDETEKSI KEMIRIPAN TEKS PADA TUGAS AKHIR MAHASISWA,” 2016.

[16] ACM Digital Library., Association for Computing Machinery. Special Interest Group on Management of Data., and Association for Computing Machinery., *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data : 2003, San Diego, California, June 09-12, 2003.* Association for Computing Machinery, 2004.

[17] N. Nurdin and A. Munthoha, “SISTEM PENDETEKSIAN KEMIRIPAN JUDUL SKRIPSI MENGGUNAKAN ALGORITMA WINNOWING,” *InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan)*, vol. 2, no. 1, pp. 90–97, Sep. 2017, doi: 10.30743/infotekjar.v2i1.165.

[18] H. Najjichah, A. Syukur, and H. Subagyo, “PENGARUH TEXT PREPROCESSING DAN KOMBINASINYA PADA PERINGKAS DOKUMEN OTOMATIS TEKS BERBAHASA INDONESIA,” 2019. [Online]. Available: http://research.

[19] I. M. Sunardi, A. Yudhana, “Perancangan Aplikasi Deteksi Plagiarisme Karya Ilmiah Menggunakan Algoritma Winnowing,” Pros. SNSebatik 2017 (Seminar Nas. Serba Inform. 2017), vol. 1, no. 1, pp. 27–32, 2017.