Fuzzy Semantic-Based String Similarity Experiments to Detect Plagiarism in Indonesian Documents

Chonan Firda Odayakana Umareta[1], Siti Mariyah[2]
[1]*STIS Polytechnic of Statistics*
Jakarta, Indonesia
[1]15.8548@stis.ac.id, [2]sitimariyah@stis.ac.id

***Abstract*— Plagiarism is a topic of concern in the world of education. One way to overcome plagiarism is to make comparisons between documents. Due to a large number of documents, extrinsic plagiarism detection frameworks are needed to make comparisons of documents in large numbers. On the other hand, there is intelligent plagiarism in which plagiarists try to hide their actions by one of them is replacing words with semantics. Therefore, this study applies an extrinsic plagiarism detection system with a Fuzzy Semantic-Based String Similarity method which is divided into three stages, namely Preprocessing, Heuristic Retrieval (HR), and Detailed Analysis (DA). In the preprocessing stage, the removal of irrelevant characters, the division of text based on sentences, stemming, tokenization, and the elimination of stopwords were performed. The search for pairs of candidate documents in the HR stage used fingerprints and Jaccard similarity. DA stage applied fuzzy semantic based-similarity. Experiments were carried out by comparing the level of document similarity between Jaccard similarity in the HR stage and fuzzy semantic-based similarity in the DA stage because both were able to produce a level of document similarity. The results show that fuzzy semantic-based similarity is better than Jaccard similarity because it can detect semantic similarities in the form of synonyms.**

Keywords—plagiarism, fuzzy, Jaccard, similarity

I. INTRODUCTION

Plagiarism is an action taken by someone in copying the contents of other people's work in whole or in part without including the source. Plagiarism is a prohibited act in the academic world and the actor will be subject to strict sanctions in accordance with applicable regulations. Based on taxonomy, plagiarism is divided into literal plagiarism and intelligent plagiarism [6]. Literal plagiarism is the most common plagiarism because it does not require much time by exact copy, near copy, and modified copy. For example, plagiarist copies and moves text from other documents by making simple changes such as adding other words, deleting several words or abbreviating sentences, dividing or combining sentences, and others. In intelligent plagiarism, plagiarists try to hide their actions by text manipulation, translation, and idea adoption. One way a plagiarist hides his behavior is to replace several words in the text that are plagiarized with synonyms so that they have a different syntax. This method is a form of intelligent plagiarism by means of the idea of adoption.

One way to prevent plagiarism is to build a detection system. The detection framework for plagiarism detection is divided into two, namely extrinsic and intrinsic [6]. Extrinsic plagiarism detection is a detection method by comparing a suspected document with a collection of source documents to determine the plagiarism portion of the source document. The final result of the detection of extrinsic plagiarism is a list of documents paired with a plagiarism chapter, which needs to be determined whether or not a violation of plagiarism is given. Intrinsic plagiarism detection is a method that only looks at a document in isolation and determines the plagiarism of the document. The purpose of detection of intrinsic plagiarism, in general, is verification of authorship to determine whether the text was actually written by the author by looking at his authorship style. Table I compares and contrasts differences between various techniques in detecting different types of plagiarism [6].

TABLE I. PLAGIARISM DETECTION METHODS AND THEIR EFFICIENCY IN DETECTING DIFFERENT PLAGIARISM TYPES

Technique	Task		Plagiarism Type								Ref
			Literal			Intelligent					
	Extrinsic	Intrinsic	Copy	Near Copy	Restructuring	Paraphrasing	Summarizing	Translating	Idea (section)	Idea (context)	
Char-Based	☑		☑	☑							[11]
Vector-Based	☑		☑	☑	☑						[12]
Syntax-Based	☑		☑	☑	☑						[13]
Semantic-Based	☑		☑	☑	☑	☑	☐				[14]
Fuzzy-Based	☑		☑	☑	☑	☑	☐				[9]
Structural-Based	☑		☑	☑	☑	☐	☐		☐	☐	[15]
Stylometric-Based		☑	☑	☑	☑						[16]
Cross-Lingual	☑		☑	☑				☑			[17]

Plagiarism can occur in various languages, one of them is Bahasa Indonesia. Because the structure of language in scientific work documents both in Bahasa and other languages has a standard writing structure and the need for further verification to determine the granting of violations of plagiarism, the plagiarism detection system suitable for academics is extrinsic. Intelligent plagiarism is serious academic dishonesty in which plagiarists try to deceive readers by changing the results of other people's contributions as if they were their own. Idea adoption is one form of intelligent plagiarism in which plagiarists use other people's ideas such as results, contributions, findings, and conclusions, without including the original source. Plagiarism ideas are classified as semantic-based meaning section-based importance, and context-based plagiarism [6]. A narrow view of the idea of plagiarism can be seen through semantic-based where the two texts have different syntax but have the same meaning.

This research aimed to build an extrinsic plagiarism detection system by applying the fuzzy semantic-based string similarity method to Indonesian documents. The detection parameters presented by [7] are not able to detect plagiarism on paper documents that we have designed plagiarism accurately in various types. So, we conducted a simulation to find the best similarity measurement parameters. One of them, we simulate the effect of the fuzzy set value of synonyms as a semantic feature on the results of plagiarism detection. Therefore, the paper is divided into five stages consisting of section 1 containing the background and objectives of this study, section 2 contains other research related to this study, section 3 contains how the plagiarism detection system works and the level of similarity is calculated, section 4 contains an explanation regarding experiments to find measurement parameters and to compare the level of similarity that has the best results in the Indonesian language test documents that were built, and section 5 contains conclusions and suggestions obtained from this research.

II. Related Work

The Fuzzy Semantic-Based String Similarity method in this research was inspired by the lab report [7] that was submitted in the PAN 2010 extrinsic plagiarism detection method competition. That plagiarism detection system is a combination of fingerprint and Jaccard similarity [2] methods used to search for candidate document pairs and fuzzy semantic-based similarity [5] with slight modifications to determine the plagiarism portion of candidate document pairs. This lab report produces recall = 0.1259, precision = 0.5761, and granularity = 3.5828 on the PAN 2010 test document in English using their best configuration.

A number of studies on the detection of plagiarism in general use cosine and Jaccard. Jaccard is one of the vector similarity matrics that can be used in measuring document similarity using vectors. Fingerprint list of unique documents is a vector of the document. The use of Fingerprint and Jaccard in performing document similarity levels has been used by [9] and [10]. Research [5] adopted the fuzzy Information Retrieval (IR) method of [9] in measuring the degree of similarity of Arabic documents. An experiment [5] was done by comparing the Boolean IR and Fuzzy IR models on the corpus that has been designed. The corpus of documents was taken from Arabic Wikipedia, and query documents or plagiarism documents were constructed manually with six scenarios according to their approach. The research showed that the Fuzzy IR model is able and better at measuring the level of similarity of documents in overcoming different writing structures and the diversity of meanings in Arabic. Research [9] used fuzzy IR in determining whether two sentences are the same or not. The determination was seen based on the level of similarity of sentences calculated using three least-frequent 4-gram approach and fuzzy-set IR in English web documents. Their experimental results showed that the fuzzy-set IR method is better than the three least-frequent 4-gram methods.

III. Extrinsic Plagiarism Detection Methods

Extrinsic Plagiarism Detection is a plagiarism detection method for comparing a suspected document with a collection of source documents. The process of detecting external or extrinsic plagiarism is divided into three stages, namely heuristic retrieval (HR), detailed analysis (DA) and post processing [4]. Before that, preprocessing is done to make the detection results better. The workflow of extrinsic plagiarism detection is illustrated in the following flowchart:

A. Preprocessing

Preprocessing is the initial stage in text processing so that analysis can be done at a later stage. Preprocessing was done on suspected documents as well as corpus documents. The process carried out in the preprocessing stage included cleansing irrelevant sentence segmentation, stemming, tokenization, and deleting stopwords. Preprocessing needs to be done to maximize the detection of plagiarism.

We omitted are all punctuation marks and other characters except sentence delimiter punctuation in the form of ".", "?", and "!". Sentence delimiter punctuations were used to do sentence segmentation. In order to change words into basic words in Indonesian documents, stemming was performed using literary stemmer [3]. Tokenization functions segmented text based on words. Then, we removed stopwords which based on source from Tala [8] as many as 758 words using an analysis of the frequency of occurrence of words.

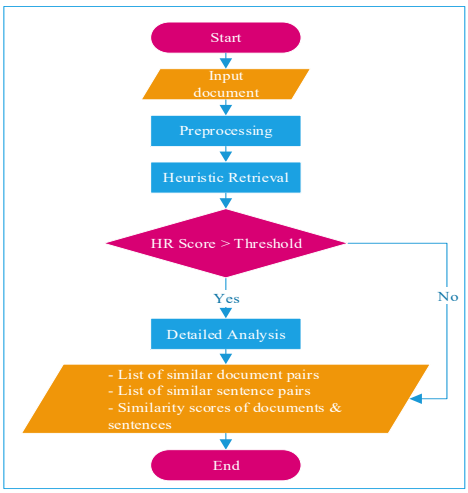

Fig. 1. Flowchart of extrinsic plagiarism detection

B. Heuristic Retrieval

Heuristic retrieval (HR) is the stage of searching for candidate pairs of suspected documents and similar source documents [4]. HR was conducted by comparing suspected documents with each source document. Documents that do not have a candidate pair were not carried out detailed analysis so as to save processing time. One of the commonly used heuristic retrieval models is fingerprinting (or shingling) [6] where documents are suspected and source documents are divided into word-k-grams based on k number of consecutive words. Word-k-grams are a representation of the k-tokens vector. Vectors can be used to measure document similarity by using vector similarity metrics (VSM). One of the VSM that can search for documents in common based on the share of a sufficient number of fingerprints is Jaccard Similarity.

Jaccard Similarity calculates the degree of similarity between documents by comparing the fingerprints of the two documents. Pairs of documents that have a Jaccard similarity more than the specified threshold were considered to be a candidate pair of similar documents and then did a detailed analysis. The pairs of documents that have a similarity level below the threshold would not be carried out detailed analysis so as to reduce computing time. Jaccard Similarity between document A and document B can be calculated using the following equation:

$$J(A,B) = \frac{|\text{fingerprint of A} \cap \text{fingerprint of B}|}{|\text{fingerprint of A} \cup \text{fingerprint of B}|} \quad (1)$$

With this, there were two parameters of plagiarism detection that need to be set to get the best results, namely the number of k-fingerprints and the threshold value for taking heuristics.

C. Detailed Analysis

Detailed analysis (DA) aims to find the most plagiarized part of the document under investigation [4]. The method used in DA was a fuzzy semantic-based similarity. In this method, the part of the plagiarism to be investigated is in the form of sentences. The list of clean sentences was first checked to filter sentences that were in accordance with Indonesian standard rules. The core structure of an Indonesian sentence is consisting of subject + predicate which can be added to the object, supplement, and/or description. The subject consists of nouns, while the predicate consists of verbs, nouns, and adjectives. Based on that, clean sentences were considered according to standard rules if they have two or more tokens which one token is a predicate and another token is the subject. Sentences that are in accordance with Indonesian language standards were processed in detailed analysis vice versa.

Fuzzy Semantic-Based Similarity is a method of measuring the degree of similarity of sentences by comparing the words in both sentences to get a sentence that is similar [7]. To get the level of similarity between two sentences (s_q, s_x) that is suspicious sentences s_q with source sentences s_x, the correlation factor of sentence relationships for each word was calculated by the formula:

$$\mu_{q,x} = 1 - \prod w_k \in s_x (1 - F_{q,k}) \quad (2)$$

Where $\mu_{q,x}$ is a correlation factor of word w_q in sentence s_q with sentence s_x, and s_x is a word in sentence w_k. $F_{q,k}$ is the fuzzy similarity between w_q and w_k which is defined by [7] as follows:

$$F_{q,k} \begin{cases} 1 & \text{if } w_k \text{ and } w_q \text{ are identical} \\ 0{,}5 & \text{If } w_k \text{ is in the synonym set of } w_q \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

Synonym is a form of language whose meaning is similar to other forms of language [19] so that fuzzy similarity between w_q and w_k which is a synonym pair of words can be considered an identical word so that it is worth 1. A list of synonyms is obtained from the Indonesian Thesaurus [18] which contains 20139 keys as entry of the word you want to search.

Correlation factor between sentences for each word w_q in the sentence s_q and the sentence s_x is then used to calculate the degree of similarity between sentences (s_q, s_x) using the equation:

$$Sim(s_q, s_x) = (\mu_{1,x} + \mu_{2,x} + \dots \quad \dots \quad \dots \quad (4)$$

Where n is the number of words in the sentence s_q. For more details, an example of calculating the degree of similarity between sentences S1 and S2 is explained in Fig. 2.
In calculating the degree of similarity of sentences between S1 and S2, the calculation was done by comparing S1 as s_q and S2 as s_x to obtain a Sim (S1, S2) = 0.625. The calculation was also done vice versa by comparing S2 as s_q and S1 as s_x to obtain a Sim (S2, S1) = 0.5. In this case Sim (S1, S2) Sim (S2, S1) because of different values. A sentence was said to be the same (EQ) if the value of the minimum sentence similarity between S1 and S2 is more than the threshold specified [5], as follows:

$$EQ(s_q, s_x) \begin{cases} 1 & \text{if } MIN(Sim(s_q, s_x), Sim(s_q, s_x)) \geq \alpha \\ 0 & \text{Otherwise} \end{cases} \quad (5)$$

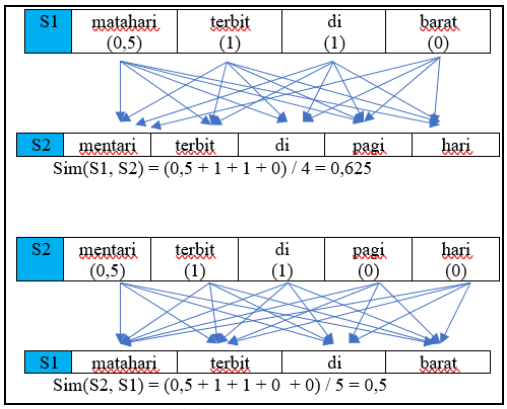

Fig. 2. Examples of different sentence pairs

To get the degree of document similarity in DA, firstly calculated the degree of similarity in all pairs (s_q, s_x) considered by EQ in (5) as follows:

$$DocSim(CDoc, QDoc) = \sum Sim(S_i, S_j) \quad \text{where} \quad EQ(S_i, S_j) = 1 \quad (6)$$

Then the similarity level of documents was obtained by calculating the Average Similarity Value (ASV) as follows:

$$ASV = DocSim(CDoc, QDoc) / N \quad (7)$$

Where N is the number of sentences in the suspected document $QDoc$. Determination of the degree of similarity of sentences for different scores between S1 and S2 and S2 with S1 does not must to choose a minimum, can be either mean or maximal. Therefore, there were two parameters of plagiarism detection for detailed analysis that need to be regulated to get the best results, namely determining the level of similarity of sentences and threshold values.

D. Evaluation

In measuring the effectiveness of a system, the two most frequent and basic measures used for the effectiveness of information retrieval were precision and recall where the IR system returns a set of query documents [1].

Precision (P) is part of the relevant accepted document.

$$P = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved}) \quad (8)$$

While the recall (R) is part of the relevant documents received.

$$R = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant}) \quad (9)$$

A measure that can hold both precision and recall values is the F-measures. Fuzzy recall and fuzzy precision were used in evaluating IR fuzzy systems [5]. Fuzzy recall is the entry level of received fuzzy set or retrieved fuzzy set (F_{RT}) in the ideal fuzzy set or ideal fuzzy set (F_{RL}), whereas fuzzy precision is the entry level of the ideal fuzzy set (F_{RL}) in the received fuzzy set (F_{RT}).

IV. Experimental Results

There were two objectives in testing the plagiarism detection system. The first was to find the configuration of the system that has the best performance for similarity value using Jaccard similarity in HR and fuzzy semantic-based similarity in DA. Secondly, a comparison of similarity values in HR and DA was carried out with the best configuration.

A. Experimental Data

Test documents were taken from 16 Indonesian documents of which eight documents were suspected or query documents (QDoc) and other eight documents were source or corpus documents (CDoc). Paper documents were suspected and source documents had different topics. Each QDoc was designed to have exactly one plagiarism against one CDoc. Plagiarism documents were made by replacing one chapter in QDoc with one chapter on a CDoc using the following test case:

1. One chapter of the QDoc1 and QDoc2 documents was a duplicate chapter of the CDoc1 and CDoc2 documents.
2. One chapter of the QDoc3 and QDoc4 documents was a closely related chapter in which all the sentences in the plagiarized chapter are entered but with words, sentences, and paragraphs that were restructured with CDoc3 and CDoc4 documents.
3. One chapter of the QDoc5 and QDoc6 documents was a chapter taken from the CDoc5 and CDoc6 documents, but each chapter included other unrelated sentences.
4. One paragraph from QDoc7 and QDoc8 was a chapter that moderates with CDoc7 and CDoc8 documents, but the words in that chapter were replaced with synonyms.

TABLE II. Experimental Result Of Heuristic Retrieval

Whole Document									
T	K = 2			K = 3			K = 4		
	P	R	F	P	R	F	P	R	F
0.1	0.125	1	0.222	0.125	1	0.222	0.125	1	0.222
0.08	0.25	1	0.4	0.25	1	0.4	0.125	1	0.222
0.06	0.5	1	0.667	0.375	1	0.545	0.25	1	0.4
0.04	0.75	1	0.857	0.625	1	0.769	0.375	1	0.545
0.02	0.875	0.875	0.875	0.75	1	0.857	0.75	1	0.857
0.01	0.875	0.467	0.609	0.75	1	0.857	0.75	1	0.857
Selected Chapter									
T	K = 2			K = 3			K = 4		
	P	R	F	P	R	F	P	R	F
1	0.286	1	0.444	0.25	1	0.4	0.25	1	0.4
0.8	0.375	1	0.545	0.375	1	0.545	0.25	1	0.4
0.6	0.625	1	0.769	0.375	1	0.545	0.375	1	0.545
0.4	0.75	1	0.857	0.625	1	0.769	0.5	1	0.667
0.2	0.75	1	0.857	0.75	1	0.857	0.75	1	0.857
0.1	0.875	1	0.933	0.75	1	0.857	0.75	0.857	0.8

B. *Experiment Results of Heuristic Retrieval Configuration*

In HR configuration experiments using Jaccard similarity, there were two points that need to be tested to get the best results, namely the number of fingerprints word-k-grams and the Threshold (T) value. The number of fingerprints tested was word-2-grams or phrases up to word-4 grams. Experimental scenario was performed by comparing the configuration of plagiarism detection in full documents or only selected chapters that were designed for plagiarism.

C. *Experiment Results of Detailed Analysis Configuration*

HR experimental results only affected the pair of documents to be carried out by the DA. HR configuration had no direct effect on DA stage performance. Therefore, experiments at the DA stage were carried out without regard to the configuration along with the pair of candidate documents received at the HR stage. DA experiments were only performed in chapters designed in each document pair.

In the DA stage, the level of similarity between sentences one and two sentences was different from the level of similarity between sentences two with one sentence. Therefore, to determine the degree of similarity of the two sentences, it was necessary to determine the value chosen whether minimum, mean, or maximum. Then, the two sentences were declared similar if the degree of similarity of the two sentences chosen was in the form of a minimum, mean, or maximum, exceeding the specified Threshold (T) value. Therefore, configuration experiments in the DA stage are conducted by comparing the best configuration in the form of a Threshold (T) value and selecting the degree of similarity in the form of minimum, mean, and maximal use of fuzzy word values of 0.5 and 1. The test documents used only focus on QDoc and CDoc pairs with selected chapters. Table III is the result of DA configuration experiments on the use of synonym fuzzy set values of 0.5.

TABLE III. EXPERIMENTAL RESULT OF DETAILED ANALYSIS

Synonym fuzzy set = 0.5									
T	MIN			MEAN			MAX		
	P	R	F	P	R	F	P	R	F
0.5	0.886	0.852	0.869	0.911	0.667	0.77	0.911	0.446	0.599
0.55	0.886	0.893	0.89	0.886	0.741	0.807	0.911	0.473	0.622
0.6	0.862	0.938	0.898	0.87	0.907	0.888	0.911	0.633	0.747
0.65	0.813	0.980	0.889	0.846	0.937	0.889	0.846	0.722	0.779
0.7	0.813	0.99	0.893	0.837	0.981	0.904	0.862	0.869	0865
Synonym fuzzy set = 1									
T	MIN			MEAN			MAX		
	P	R	F	P	R	F	P	R	F
0.5	0.902	0.707	0.793	0.919	0.642	0.756	0.927	0.429	0.586
0.55	0.886	0.893	0.89	0.894	0.775	0.830	0.927	0.452	0.608
0.6	0.87	0.947	0.907	0.878	0.908	0.893	0.902	0.624	0.738
0.65	0.837	0.963	0.896	0.862	0.922	0.891	0.87	0.699	0.775
0.7	0.821	1	0.902	0.846	0.981	0.908	0.846	0.839	0.842

The best configuration on the use of fuzzy sets of synonym words of 0.5 and 1 could be obtained using the configuration of the choice of sentence similarity level in the form of mean and threshold value of 0.7. Overall, the best configuration for a fuzzy set of synonym words of 1 was slightly better than using a fuzzy set of synonym words of 0.5. For more details, Fig. 3 is

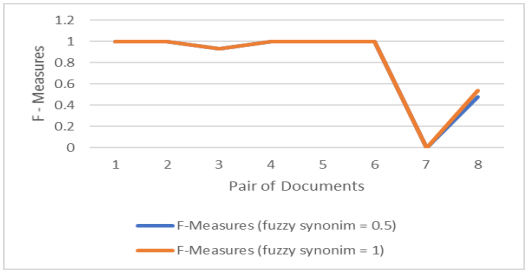

Fig. 3. Performance Evaluation

a performance graph on the fuzzy set for each pair of documents.

Based on Fig. 3, the use of fuzzy synonym set of 1 had better performance than 0.5. The difference lies in the 8th pair of documents where the F-Measures on the synonym fuzzy set are 1 higher than the fuzzy synonym set of 0.5. The 8th plagiarism document pair was a document pair which was simulated plagiarism by replacing each word into its synonym. Therefore, it can be concluded that the use of fuzzy synonym sets of 1 is better than the use of fuzzy synonym sets of 0.5 in detecting document plagiarism, especially in the form of plagiarism by replacing word synonyms.

D. *Comparison of Document Similarity*

The HR stage and DA stage produce different degree of document similarity with their respective calculation methods. The HR stage achieved the degree of similarity of documents using Jaccard similarity, whereas the DA stage calculated the degree of document similarity using Average Similarity Value (ASV). The degree of similarity of documents using ASV can only be obtained if the DA is performed on pairs of candidate documents that have a degree of similarity of documents using Jaccard similarity more than the threshold specified at the HR stage. The HR and DA stages also have their respective configurations that affect the results of calculating the degree of similarity. The following is a comparison chart of the degree of similarity of documents with Jaccard similarity using k = 2, and ASV using fuzzy synonym sets of 0.5 and 1.

ASV yields a higher degree of similarity than Jaccard, and also ASV with a fuzzy set of synonyms valued at 1 yield a higher level than 0.5. ASV is better than Jaccard in the 3rd and 4th pairs of documents that are simulated plagiarism by restructuring words, sentences, and paragraphs, and in the 8th pair of documents that are simulated plagiarism by replacing each word using its synonyms. ASV by using a fuzzy set of synonyms of 1 is better than 0.5 in the 8th pair of documents simulated plagiarism by replacing each word using its synonym.

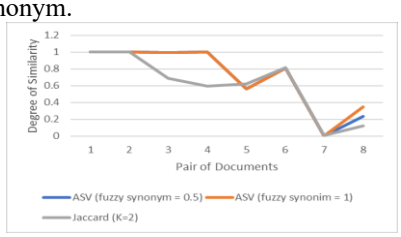

Fig. 4. Degree of Similarity Evaluation (Jaccard vs ASV)

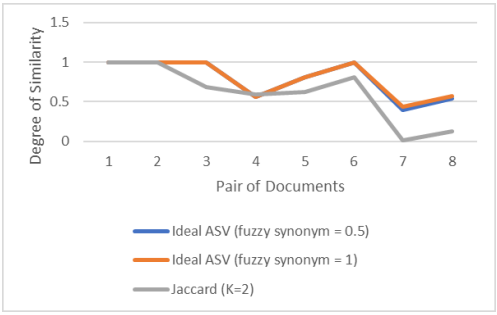

Fig. 5. Degree of similarity evaluation (Jaccard vs Ideal ASV)

In the 7th pair of documents, which were also synonymous plagiarism, ASV which should be able to detect the use of synonyms did not produce a degree of similarity. It was caused by all the pairs of sentences designed for plagiarism has a level of similarity below the best-determined threshold configuration so it was not included in ASV calculation. Therefore, an ideal ASV calculation was performed to see the pair of sentences designed by plagiarism by ignoring the threshold since it was not detected as plagiarism (below the threshold).

After calculating the ideal ASV, the degree of similarity of the 7th pair of documents could be found. The degree of similarity was less than 0.5 and was below the best threshold obtained, which was 0.7. If the threshold was forced down to detect this, and then the performance would decrease because other sentences that were not plagiarism would also be detected as plagiarism.

V. Conclusion

Intelligent plagiarism is serious academic dishonesty in which plagiarists try to deceive readers by changing the results of other people's contributions like the research ideas such as results, contributions, findings, and conclusions, without including the original source as if they were their own. One way a plagiarist hides his behavior is to replace several words in the text that are plagiarized with synonyms, and also the large number of documents that must be investigated makes it difficult for investigators. Therefore, this research aimed to build an extrinsic plagiarism detection system by applying the fuzzy semantic-based string similarity method to Indonesian documents. The simulation is done by finding the best similarity measurement parameters that can produce the best performance in the test document. This research shows that the best parameter for the HR stage is k-fingerprints = 2 and threshold value of 0.1 with F-measures of 0.933. The best parameter for the DA stage among others the choice of sentence similarity level in the form of mean and threshold value of 0.7 with F-measures of 0.981. This detection system can also be used for various language. The inadequate use of Jaccard similarity in detecting synonymous plagiarism at the HR stage can pass plagiarism documents that way from the DA stage which is capable of detecting synonym plagiarism. So for the future work, we need a plagiarism detection method that is able to process the entire contents of the document in quick time and be able to detect the use of the word synonym in the plagiarism document to be used in finding candidate pairs of documents that are suspected of plagiarism.

References

[1] C. D. Manning, P. Raghavan, and H. Schutze, “Evaluation in information retrieval: Evaluation of unranked retrieval sets,” in Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 154–158.

[2] C. D. Manning, P. Raghavan, and H. Schutze, “Web search basics: Near-duplicates and shingling,” in Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 437–442.

[3] Librian, A. High quality stemmer library for Indonesian Language (Bahasa), GitHub, 2017. Available at: https://github.com/sastrawi (Accessed: 14 August 2019).

[4] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., & Rosso, P. Overview of the 1st International Competition on Plagiarism Detection. San Sebastian: Spanish Society for Natural Language Processing (SEPLN), 2009.

[5] S. Alzahrani and N. Salim, “On the use of fuzzy information retrieval for gauging similarity of arabic documents,” in Proc. 2nd Int. Conf. Appl. Digital Inf. Web Technol., 2009, pp. 539–544.

[6] S. M Alzahrani, N. Salim, and A. Abraham. “Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods.” IEEE Systems, Man, and Cybernetics Society, 2011: 1-7.

[7] S. Alzahrani and N. Salim, “Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF’10.” presented at the 4th Int. Workshop PAN-10. Padua, Italy, 2010.

[8] Tala, F. Z. “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia,” M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteit van Amsterdam, The Netherlands, 2003.

[9] Yerra, R. and Y.-K. Ng, “A Sentence-Based Copy Detection Approach for Web Documents”, Fuzzy Systems and Knowledge Discovery, 2005, pp. 557-570.

[10] J. Kasprzak, M. Brandejs, and M. K˘ripac, “Finding plagiarism by evaluating document similarities,” in Proc. SEPLN, Donostia, Spain, pp. 24–28.

[11] C. Grozea, C. Gehl, and M. Popescu, “ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection,” in Proc. SEPLN, Donostia, Spain, 2012, pp. 10–18.

[12] A. Barron-Cede ´ no, C. Basile, M. Degli Esposti, and P. Rosso, “Word length n-Grams for text re-use detection,” in Computational Linguistics and Intelligent Text Processing, 2010, pp. 687–699.

[13] M. Elhadi and A. Al-Tobi, “Use of text syntactical structures in detection of document duplicates,” in Proc. 3rd Int. Conf. Digital Inf. Manage., London, U.K., 2008, pp. 520–525.

[14] Y. Li, D. McLean, Z. A. Bandar, J. D. O’Shea, and K. Crockett, “Sentence similarity based on semantic nets and corpus statistics,” IEEE Trans. Knowl. Data Eng., vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[15] H. Zhang and T. W. S. Chow, “A coarse-to-fine framework to efficiently thwart plagiarism,” Pattern Recog., vol. 44, pp. 471–487, 2011

[16] E. Stamatatos, “A survey of modern authorship attribution methods,” J. Amer. Soc. Inf. Sci. Technol., vol. 60, pp. 538–556, 2009.

[17] M. Potthast, A. Barron-Cede ´ no, B. Stein, and P. Rosso, “Cross-language plagiarism detection,” Language Resources & Evaluation, pp. 1–18, 2010.

[18] Pusat Bahasa Departemen Pendidikan Nasional. Tesaurus Bahasa Indonesia Pusat Bahasa. Jakarta: Departemen Pendidikan Nasional, 2008.

[19] Pusat Bahasa Departemen Pendidikan Nasional. Kamus Bahasa Indonesia. Jakarta: Departemen Pendidikan Nasional, 2008.