ISSN 2622-626X

A review of detection plagiarism in indonesian language

Ida Widaningrum [a,1,*], Dyah Mustikasari [a,2], Rizal Arifin [b,3] , Sugianti[a,4]

a Informatics Universitas Muhammadiyah Ponorogo, Jl. Budi Utomo N0. 10, Ponorogo and 63471, Indonesia
b mechanical engineering Universitas Muhammadiyah Ponorogo, Jl. Budi Utomo N0. 10, Ponorogo and 63471, Indonesia
[1] iwidaningrum.as@gmail.com*; [2] dyah.mustikasari@gmail.com; [3] rizal.arifin@gmail.com; [4] giantisugianti@gmail.com

ARTICLE INFO

Article history
Received 2018-09-14
Revised 2018-11-10
Accepted 2018-12-15

Keywords
plagiarism
creative ideas or ideas
Indonesian

ABSTRACT

Plagiarism is the act of copying the work of another person in the form of writing, ideas, creative ideas or other without including the source of the work or idea. This action is of course very disrespectful, violates the code of ethics and is opposed by all parties, both by scientists and government. This happens because the use of the internet provides unlimited information services. Many studies have been carried out, raising the theme of this plagiarism. This article will review how far the plagiarism research has been done on Indonesian writing. By knowing the development of plagiarism research, further research will have better sustainability.

1. Introduction

Plagiarism is the disrespectful act of copying the work of another person without mentioning the original source of the work. It becomes a serious problem in scientific writing. This act violates the ethics of academic culture that always gives the acknowledgment to the authors of the cited articles. The scientists, the academicians and the government of Indonesia have been seriously against the act of plagiarism during the recent years. The government of Indonesia has issued the law to protect the intellectual properties [1] and to prevent the plagiarism in the higher education institutions [2] [3]. Undang-Undang Nomor 19 Tahun 2002 Tentang Hak Cipta [1], Undang-Undang Nomor 20 Tahun 2003 Tentang Sistem Pendidikan Nasional [2], and Permendiknas Nomor 17 Tahun 2010 Tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi [3]. Furthermore, the number of researches aiming to develop the plagiarism detection tools is increasing. In this article, we review the research progress on the Indonesian language plagiarism detection theme. This paper will give the knowledge to the novel and senior researchers that are interested in the topic of plagiarism detection for the Indonesian language.

2. Plagiarism

Definition

According to Indonesian dictionary [4] [5], Plagiarism is taking a work (ideas etc.) of others and making it (ideas etc.) as his own (the plagiarist), for example publishing a work of other people under his own name. According to Merriam-Webster dictionary [6], plagiarism is the act of stealing and recognizing ideas or the work of others as their own (the plagiarist). Sometimes plagiarism is committed by mistakes due to lack of knowledge in citing works. Basically, plagiarism is the act of taking ideas, thoughts, or works of others without any references [7], [8] whether it is intentionally or not. To avoid plagiarism, methods were introduced to detect to what extent the articles we wrote contained plagiarism elements, one of them is to detect plagiarism in Indonesian articles. Detection on articles delivered in Indonesian are treated differently from English articles due to the different

http://dx.doi.org/10.33292/ijarlit.v1i2.27

language structure. Therefore, we will study what methods have been used and have been investigated to detect the plagiarism in Indonesian articles.

Types of Plagiarism

There are several types or characters of plagiarism, namely plagiarism by copying without reordering the initial sentence or directly copying. It can be copying by rearrange a few sentences by changing the words order, changing the active sentence to passive, replacing the words with other words having same meaning, direct translation or repetitive translation, quoting the structure of writing, and etc. Plagiarism consists of literal plagiarism and intelligent plagiarism [8]. Literal plagiarism is copying the whole, parts or modifying the text without mentioning the original author in the references. While the intelligent plagiarism is changing the contribution of the original author as if the materials belong to the plagiarist, trying to hide, obscuring and rearranging the original work including manipulation of text, translation, and ideas. While [9], classified plagiarism into two, namely textual plagiarism and source code plagiarism. Textual plagiarism includes copy paste / clone plagiarism, paraphrasing consists of simple paraphrasing and mosaic / hybrid / patchwork paraphrasing, metaphor, idea, self-recycled, illegitimate source, and retweet plagiarism. Source code plagiarism consists of manipulation from vicinity plagiarism, reordering structure, no change plagiarism, and language switching plagiarism.

2. Plagiarism Detection Methods

To check whether or not a document contains plagiarism, it is inputted into a system to be analyzed in which the system had contained documents that could be used as the comparison. Then, it was analyzed by the existing method and finally the suspected part of plagiarism could be determined. The next step was confirmation and the last was investigation. The illustration of plagiarism detection system is as follows:

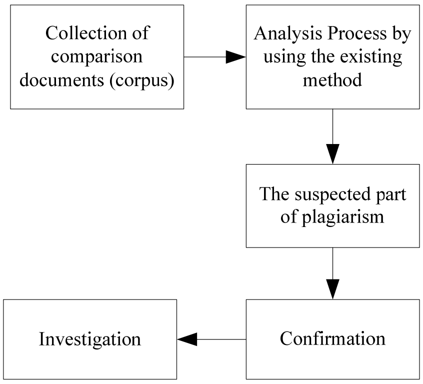
Collection of comparison documents (corpus)
Analysis Process by using the existing method
The suspected part of plagiarism
Confirmation
Investigation

Figure 1. Plagiarism Detection Process Flowchart

There are a lot of classifications to detect plagiarism that have been proposed, namely:

a. The three approaches, namely natural language approach, index structure approach, external plagiarism detection and clustering. Natural language approach is further divided into grammar, semantic, and semantic hybrid methods [7]

b. Plagiarism detection is divided into textual plagiarism, citation based and shape-based PD for flowchart. Textual plagiarism is divided into grammar based method and external plagiarism detection method [10].

c. Detection based on the tasks and techniques. Detection based on the task is divided into extrinsic and intrinsic plagiarism detection [11]. Whereas based on the technique, it consists of character-based method which is further divided into; fingerprint, string similarity, structural-based method, classification and cluster-based method, syntax-based method, cross language-based method, semantic-based method, and citation-based method.

d. [9] Plagiarism detection is divided into monolingual plagiarism and cross-lingual plagiarism detection. Monolingual is further divided into intrinsic plagiarism and extrinsic plagiarism. It divides the method into character, vector, syntax, semantics, fuzzy, structure, stylometric, cross-lingual, semantic hybrid grammar, classification and clustering, and finally citation-based methods.

e. While [8] classified detection methods into plagiarism detection based on:

1) The Task; Extrinsic and Intrinsic.
2) Language; Monolingual and cross-lingual.
3) Textual features; for extrinsic plagiarism detection, intrinsic plagiarism detection, and cross-lingual plagiarism detection.

Textual features for extrinsic plagiarism detection;

a. Lexical Feature: working on the character or word level, commonly called as fingerprint or shingles. For example, Character-based n-gram (CNG), Word-based n-gram (WNG).

b. Syntactic Features: Manifestation of part of speech (POS) from phrases and words in different sentences. Basics of POS tag include verbs, nouns, pronouns, verbs, adverbs, prepositions, conjunctions, and interjections. POS tagging tasks are marking words in statements that match on certain POS tags.

c. Semantic features: quantifying the use of word classes, synonyms, antonyms, hypernyms, and hyponyms. The use of thesaurus dictionary and lexical database will provide more insight into the meaning of semantic texts. Working together with POS tagging, it can help the detection.

d. Structural features: reflecting text organization and capture more comprehensive semantics. Structural features may characterize the documents as headers, sections, sub-sections, paragraphs, sentences, etc.

Textual features for intrinsic plagiarism detection based or Stylometry features are based on the writing style of each author, considering the aspects of style including;

a. Text statistics through various lexical features, operating on character or word levels;

b. Syntactic features, working at sentence level, measuring the use of word classes, and / or sorting sentences into parts of speech;

c. Semantic features, measuring the use of synonyms, functional words, and / or semantic dependencies; and

d. Application-specific features, reflecting text organization, content-specific keywords, and / or other language-specific features

From the description of several research above, it can be concluded that we may return to the outline for this plagiarism detection method which is divided into two, extrinsic and intrinsic plagiarism detection.

Extrinsic Plagiarism Detection

Extrinsic plagiarism detection is a method of comparing query documents or suspicious documents with the comparative documents. There are a lot of research developed, which [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23].

Intrinsic Plagiarism Detection

Intrinsic plagiarism detection tends to be more focused on the writing style of the author. Every author has a certain style in writing. In this intrinsic plagiarism detection, writing style is examined whether the article has same writing style from beginning to the end. If it has differences in certain parts, it means that the article was taken from others sources without mentioned it in the references [24], [25] [26] [27] [28] [29] [30].

Various research that have been collected and identified showed that the algorithm applied for plagiarism detection in Indonesian documents is quite large. If it is examined, this algorithm is divided into two groups. The first group is based on grammar and the other is based on semantics. Grammar-based research uses Rabin Karp algorithm [31] [32]

3. Discussion and Result

Indonesian Language Plagiarism Detection

Many research on detection plagiarism in Indonesian documents have been carried out. The focus is on the effectiveness of algorithms for detecting plagiarism. Of the many algorithms studied, Rabin-Karp became the most widely applied [31-45]. According to [31] this algorithm will be effectively used to search for multiple pattern rather than single pattern. It is used with Dice similarity coefficient to calculate the percentage [33]. [34]carried out detection using Rabin-Karp algorithm with rolling hash method and implemented into a program or application to determine accuracy level with a percentage value. Based on the analysis of plagiarism detecting process using raibin-karp algorithm by rolling hash method, it can read the character in the form of letters, symbols such as dots (.), Commas (,), etc.

The use of Rabin-Karp algorithm [35] as a string matching algorithm by multiple search method can detect plagiarism based on similarity level. Mathematical approach is used to get the ideal k-gram and modulo values. Mode function to calculate k-grm value in parsing process and the largest prime number approach to calculate modulo in hashing process. The test results on 7 data with 2 variants and 3 treatments with and without stemming indicate the mode function in similarity testing shows that mean function without stemming is greater than that of with stemming, and applies for median function as well. The biggest prime number approach effectively generates ideal prime number or modulo value that can be used in hashing process and eliminates quadratic occurrence when string matching uses Rabin-Karp algorithm.

Another research on Rabin-Karp's main parameters, which are k-gram, base (number of word characters) and modulo (determined prime number) [36], stated that right combination determination will produce good accuracy. Decreasing k-gram can increase plagiarism accuracy, as well as value of base and modulo that will increase the accuracy.

Rabin-Karp was compared to others algorithm such are Knut-Morris-Pratt Algorithm [33], Winnowing Algorithm [37] and Levenshtein Distance [38]. The result of comparation by [39] was that each of algorithm has its strengths and weaknesses. Here, it is said that Knut-Moris-Pratt is more suitable for string search cases in general, whereas Rabin-Karp is more suitable for searching strings with long patterns, multi patterns, imperfectly repeated patterns in text fields, strings with patterns and text fields with relatively the same lengths or strings with different text field to the first character with its pattern.

In another comparation research between Rabin-Karp and Winnowing [37], it was said that. Winnowing Algorithm approach is better than Rabin Karp algorithm since it produces a smaller percentage rate and faster processing time. Based on result of the test on winnowing and Rabin Karp algorithm approaches, it can be seen that smallest similarity of thesis title using Winnowing algorithm approach is on the 8th test with n-gram = 9 and window = 3, 0.0257 with the smallest similarity of 32.6%.

[38] performed another comparation of Rabin-Karp and Levenshtein Distance algorithm. The conducted process in the system is preprocessing on each algorithm then calculating similarity level using Rabin-Karp and Levenshtein Distance algorithm. So that it obtains the result of two algorithms existing in the system to be analyzed their comparison results in the form of graphs. In the testing, the researcher uses text documents with .pdf format in Indonesian. With the same data set, Rabin-Karp took more time than Levenshtein Distance algorithm.

Rabin-Karp was also used along with Nazief-andriani algorithm for stemming process [40]. Nazief-andriani algorithm is stemming algorithm for Bahasa Indonesia. The accuracy of Nazief-andriani stemming algorithm is strongly affected by the dictionary comprehensiveness of root words. This study stated that the more comprehensive, the more accurate the stemming result. The last step after string matching is calculating the similarity [40]. The influence of this stemming algorithm was also applied along with Fingerprint Matching [41]. The result is the use of stemming with stopword removal can detect 42% better than using stemming alone by 31% or without pre-processing by 34% when applying bigram.

Another algorithm used to detect plagiarism in Indonesia languge is Winnowing [42]. It is used to detect text files in one-to-one, one-to-many, even many-to-many. In detecting similarities among text files, Winnowing Algorithm is used. This algorithm serves to analyze fingerprinting document which converts text into a set of hash values. This application uses a database in XML format files to run without requiring complicated database configurations. [43]also performed plagiarism detection by Winnowing algorithm. Using fingerprint with Phrase-based technique that breaks the text document into biword tokens. Then it is encrypted into an MD5 value, so it has the same hash value and can be used as a text document fingerprint. Therefore, phrases checking for each document can be conducted and then save it in an array. This algorithm is also compared to Fingerprint by [44]. The study yield that fingerprint is better than Winnowing with 92.8% while Winnowing only got 91.8% in perfomance. At the level-of-relevance, Winnowing scored on 37.1% term-correlation while fingerprint got 33.6%.

For detection based on semantic, [45] employed LSA (Latent Semantic Index). Semantic analysis is performed using WordNet in Indonesian as the source. Several calculation to measure the degree of similarity such as Jaccard, Pearson, and Euclidean distance were used and compared by using a combination of hashing function and semantic analysis. The calculation result indicates that both coefficient calculations provide a stable condition after a number of hash functions. Pearson coefficient provides maximum result because it can detect the similar data but uses different value accurately as indicated for non identical documents detection. Meanwhile, for identical documents, Jaccard shows a better accuracy than Pearson. However, to achieve the right results, the number of hash functions on Jaccard should be increased since the result tends to give a relatively higher error than Pearson. Another study of LSA in Indonesian plagiarism is performed by [46]. This used LSA in Heuristic and Detailed Analysis Component. The result is LSA better than VSM in performance, particularly on intelligence plagiarism.

The Knut-Moris-Pratt algorithm is also used by [47] to detect develop PlagON, a similarity indication engine for document written in Bahasa Indonesia. This engine used Nazief-andriani for stemming and Dice Coefficient for calculating the similarity.

A new approach in plagiarism detection system called Citation-based Plagiarism Detection is tackled by [48] to detect plagiarism in Indonesian text. This approach analyzes quotations so that it allows duplicate and plagiarism to be detected even if the document has been paraphrased or translated due to the same quotation position. Kang algorithm is one method to detect suspected plagiarism based on text comparison. This method has the ability to detect documents similarity percentage and possible plagiarism types by checking the document sentence per sentence compared to other documents sentence per sentence. Citation-based Plagiarism Detection (CbPD) and Kang algorithm can be a solution to improve the efficiency in plagiarism detection without sacrifices the accuracy. In addition, some adjustments are needed in the plagiarism detection system to perform the concept of CbPD and Kang algorithm (CbPD-Kang). Several required mechanisms in this approach are similarity detection in references written in different formats and document citation pattern. The method used in determining document priority is Cosine Similarity.

The similarity measure is also compared [49]. This is performed along with the Rabin-Karp algorithm using Indonesian texts. The similarity measure compared is Dice, Cosine, and Jaccard coefficients. Four documents which have a variety of similarities are used. The result indicates that

Cosine similarity provides better performance compared to Dice and Jaccard coefficients. This model can be used as an alternative type of n-gram process statistical algorithm.

Generally, research on plagiarism detection that has been carried out so far around the use of preprocessing algorithms and stemming. The algorithm used by Rabin Karp, Knut-Morris-Pratt and Winnowing or a combination of them. In addition, there are also those who use LSA

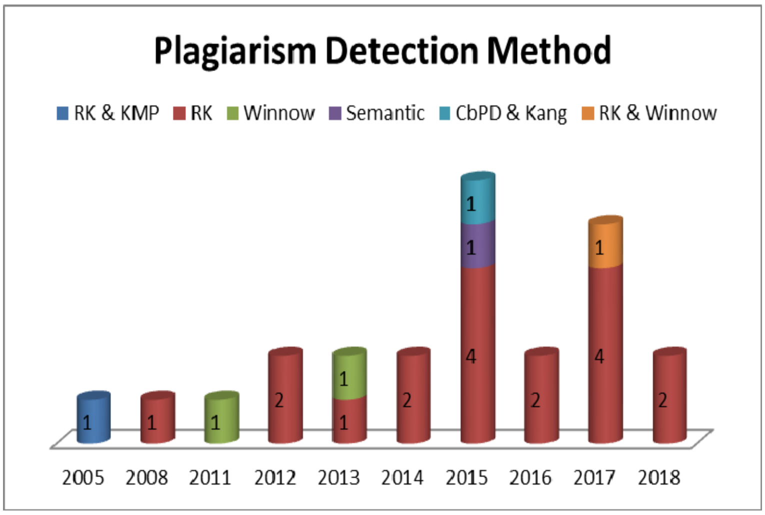

Figure 2. Plagiarism detection methods often used for Indonesian plagiarism detection

From the picture it appears, studies that use the Rabin-Karp method are more as described previously. While other studies that might be suitable for the Indonesian language based detection cases are still lacking. Thus, it should be tried and examined whether other methods are also suitable for the case of Indonesian.

Detection Plagiarism

Various collected and identified studies indicate that there are quite large algorithms applied for plagiarism detection in Indonesian documents. If it is analyzed, this algorithm is divided into two methods.

1. Grammar-Based Plagiarism Detection

Grammar is a detection method based on the sequence of words in a sentence. This method detects the string order in a document. In this method, the algorithm is suitable to detect a full (copy-paste) text without any modification or little modification.

2. Semantic-Based Plagiarism Detection

It focuses on detecting similarities between documents by using the vector space model. This method can specify and calculate the redundancy of words in a document as well, then uses fingerprints for each document and matches them with fingerprints from other documents and identifies the similarities. The semantic-based method is suitable for non-partial plagiarism. It uses vector space to match documents. However, if the document has been plagiarized, it will not generate a good result, and therefore, it is considered as a limitation of this method, as it is difficult to search the copied texts in the original document.

The methods for detecting plagiarism that can be used are very diverse including the ones illustrated in Figure 3:

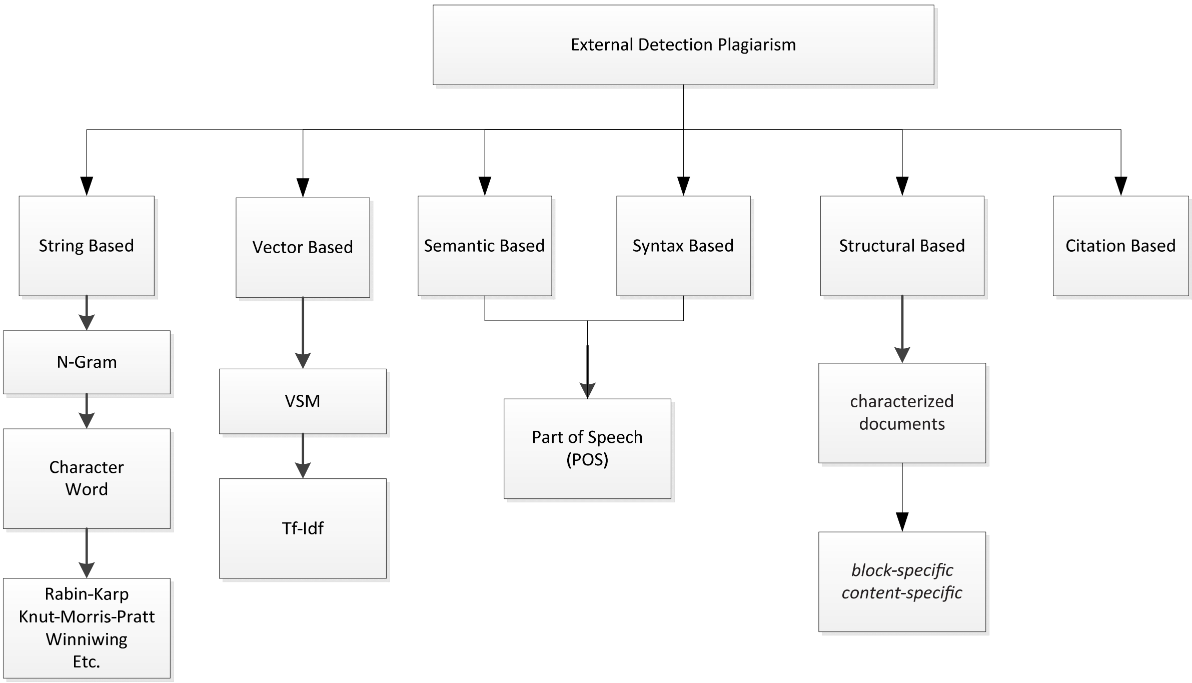

Figure 3. Plagiarism detection method.

String based is the simplest comparison, where compared are characters or words. N-grams are groups of words or characters formed from documents [12], [13], [50]. Vector based, uses syntax and lexical features to represent documents in vector space [18] [51]. The vector based model also uses the weighting term frequency-inverse document frequency (tf-idf) and the term frequency-inverse sentence frequency (tf-isf) [52]. Semantic based and syntax based, the document extracted can be a sentence, phrase or based on POS (Part Of Speech). [53]weighting with TF-ISF and POS, can actually increase precision when compared to POS. Then this was developed by [17] by using fuzzy semantics. Structural based, displays organization from text and uses more semantic documents. Documents are described as a collection of paragraphs or sections, things are used to extract information from documents [54]. Structural-based is divided into block-specific and content-specific. Block-specific structured features, are used to describe a collection of web documents as blocks, namely, paragraphs [55].

Of all the studies carried out, no one has mentioned whether the document that is the comparison (reference) must use language that is in accordance with the documents that will be examined in terms of similarity (suspect). That is, if the suspect documents speak Indonesian, references are only Indonesian language documents. Thus when comparing similarities, the similarity is only for Indonesian documents only. Whereas, as we know, many other documents use languages other than Indonesian. And maybe intentionally or not the quote is done on another document. So it needs an additional stage where we have to adjust the language of both parties, namely suspect and reference or there is a cross-language process. Generally described as follows:

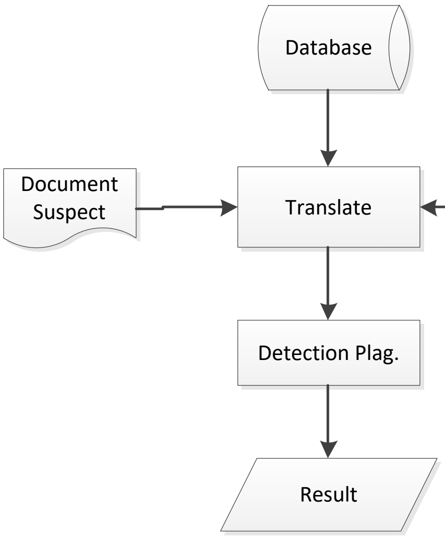

Figure 4. General architecture of plagiarism detection

4. Conclusion

From the results of review, it was seen that there had been a lot of research done but not extensive. The research was carried out only on the application of algorithms, and not all algorithms were tried and compared. Because the algorithms used are few, we cannot take the conclusions that certain algorithms are in accordance with the structure of Indonesian. Likewise when comparing the use of an algorithm, it cannot be concluded because the use of the test components is different. For further research, it is recommended to make comparisons of algorithms with the same test components. The differences will be clearly seen from one another (advantages and disadvantages).

In addition, because the documents needed as a comparison are not only Indonesian language documents or only one language, then the language matching phase is needed from the suspect document with a comparative document (reference document). This stage is called auto translation.

Acknowledgment

The author is grateful for the support of all parties who have provided funds, suggestions and contributions to their thoughts.

References

[1] Pemerintah Republik Indonesia, “Undang-Undang Republik Indonesia Nomor 19 Tahun 2002 Tentang Hak Cipta,” 2002.

[2] Pemerintah Republik Indonesia, “Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 Tentang Sistem Pendidikan Nasional,” *Dep. Pendidik. Nas.*, pp. 1–33, 2003.

[3] Kementerian Pendidikan Nasional Indonesia, “Peraturan Menteri Pendidikan Nasional Republik Indonesia tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi Nomor 17 Tahun 2010.” 2010.

[4] Http://www.kbbionline.com/arti/kbbi, “Kamus Besar Bahasa Indonesia Online,” *diakses tgl 26 September 2018*, 2018. .

[5] Https://kbbi.kemdikbud.go.id, “Kamus Besar Bahasa Indonesia Online,” *diakses tanggal 26 September 2018*. .

[6] Https://www.merriam-webster.com/dictionary, “merriam-webster dictinary online,” *diakses tanggal 26 Sept. 2018.*

[7] A. M. E. T. Ali, H. M. D. Abdulla, and V. Snasel, “Survey of Plagiarism Detection Methods,” *2011 Fifth Asia Model. Symp.*, pp. 39–42, 2011.

[8] S. M. Alzahrani, N. Salim, and A. Abraham, “Understanding plagiarism linguistic patterns, textual

features, and detection methods," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 2, pp. 133–149, 2012.

[9] H. A. Chowdhury and D. K. Bhattacharyya, "Plagiarism: Taxonomy, Tools and Detection Techniques," *19th Natl. Conv. Knowledge, Libr. Inf. Netw. (NACLIN 2016)*, no. 1, 2016.

[10] S. A. Hiremath and M. S. Otari, "Plagiarism Detection-Different Methods and Their Analysis: Review," *Int. J. Innov. Res. Adv. Eng.*, vol. 1, no. 7, pp. 2349–2163, 2014.

[11] A. S. Hamza Osman Naomie; Abuobieda, Albaraa, "Survey of Text Plagiarism Detection," *Comput. Eng. Appl. J.*, vol. 1, no. Vol 1, No 1: June 2012, pp. 37–45, 2012.

[12] C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," *CEUR Workshop Proc.*, vol. 502, pp. 10–18, 2009.

[13] C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, and M. D. Esposti, "A plagiarism detection procedure in three steps: Selection, matches and squares," *CEUR Workshop Proc.*, vol. 502, pp. 19–23, 2009.

[14] A. Barrón-Cedeño and P. Rosso, "On automatic plagiarism detection based on n-grams comparison," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5478 LNCS, pp. 696–700, 2009.

[15] M. Elhadi and A. Al-Tobi, "Use of text syntactical structures in detection of document duplicates," *3rd Int. Conf. Digit. Inf. Manag. ICDIM 2008*, pp. 520–525, 2008.

[16] J. Koberstein and Y.-K. Ng, "Using Word Clusters to Detect Similar Web Documents," pp. 215–228, 2006.

[17] S. Alzahrani and N. Salim, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF 2010," *CEUR Workshop Proc.*, vol. 1176, 2010.

[18] V. K and D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," *J. Eng. Sci. Technol. Rev.*, vol. 9, no. 4, pp. 150–164, 2016.

[19] S. S. Dharani, J. Ganesh, R. Ieshwarya, and M. Sureka, "Extrinsic Plagiarism Detection System for Semantic Replication in Medline," vol. 4, no. 11, pp. 45–50, 2016.

[20] Z. F. Alfikri and A. Purwarianti, "Detailed Analysis of Extrinsic Plagiarism Detection System Using Machine Learning Approach (Naive Bayes and SVM)," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 12, no. 11, pp. 7884–7894, 2014.

[21] M. Alsallal, R. Iqbal, S. Amin, A. James, and V. Palade, "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection," *Proc. - 2016 9th Int. Conf. Dev. eSystems Eng. DeSE 2016*, pp. 203–208, 2017.

[22] A. Magooda, A. Mahgoub, M. Rashwan, M. Fayek, and H. Raafa, "RDI System for Extrinsic Plagiarism Detection (RDI_RED)," *Work. Notes PAN-AraPlagDet FIRE 2015*, pp. 129–131, 2015.

[23] R. Naseem and S. Kurian, "Extrinsic Plagiarism Detection in Text Combining Vector Space Model and Fuzzy Semantic Similarity Scheme," *IRACST – Int. J. Adv. Comput. Eng. Appl.*, vol. 2, no. 6, pp. 2319–281, 2013.

[24] B. Stein and S. Meyer Zu Eissen, "Intrinsic plagiarism analysis with meta learning," *CEUR Workshop Proc.*, vol. 276, pp. 45–50, 2007.

[25] B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 63–82, 2011.

[26] E. Stamatatos, "Intrinsic Plagiarism Detection Using Character n -gram Profiles," 2006.

[27] S. Meyer and B. Stein, "LNCS 3936 - Intrinsic Plagiarism Detection," *Advances*, pp. 565–569, 2006.

[28] S. Meyer zu Eissen, B. Stein, and M. Kulig, "Plagiarism Detection Without Reference Collections," pp. 359–366, 2007.

[29] M. Kuznetsov, A. Motrenko, R. Kuznetsova, and V. Strijov, "Methods for intrinsic plagiarism detection and author diarization," *CEUR Workshop Proc.*, vol. 1609, pp. 912–919, 2016.

[30] L. Seaward and S. Matwin, “Intrinsic Plagiarism Detection using Complexity Analysis,” *Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. SEPLN 2009 Work. Uncovering Plagiarism, Authorship, Soc. Softw. Misuse (PAN 2009)*, pp. 56–61, 2009.

[31] N. Baedlowi, D. A. Adam, and L. Ilmu, “String Matching dengan Menggunakan Algoritma Rabin Karp,” pp. 1–3, 2008.

[32] Salmuasih and A. Sunyoto, “Implementasi Algoritma Rabin Karp untuk Pendeteksian Plagiat Dokumen Teks Menggunakan Konsep Similarity,” *Semin. Nas. Apl. Teknol. Inf. 2013*, pp. 23–28, 2013.

[33] D. A. Putra, H. Sujaini, and H. S. Pratiwi, “Implementasi Algoritma Rabin-Karp untuk Membantu Pendeteksian Plagiat pada Karya Ilmiah,” *J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1–9, 2015.

[34] J. Priambodo, “PENDETEKSIAN PLAGIARISME MENGGUNAKAN ALGORITMA RABIN-KARP DENGAN METODE ROLLING HASH,” *J. Inform. Univ. Pamulang*, vol. 3, no. 1, pp. 39–45, 2018.

[35] Y. T. Lede, P.A.R.L., Fanggidae, A. dan Polly, “Implementasi Algoritma Rabin-Karp Untuk Mendeteksi Dugaan Plagiarisme Berdasarkan Tingkat Kemiripan Kata Pada Dokumen Teks,” *J. Komput. Inform.*, vol. 2, no. 1, pp. 50–64, 2014.

[36] A. Putera Utama Siahaan and Sugianto, “Analisis k-gram, basis dan modulo rabin-karp sebagai penentu akurasi persentase kemiripan dokumen,” in *SENASPRO 2017 | Seminar Nasional dan Gelar Produk*, 2017, pp. 198–206.

[37] N. Alamsyah, “Perbandingan Algoritma Winnowing Dengan Algoritma Rabin Karp Untuk Mendeteksi Plagiarisme Pada Kemiripan Teks Judul Skripsi,” *Technologia*, vol. 8, no. 3, pp. 124–134, 2017.

[38] A. H. Purba and Z. Situmorang, “Analisis Perbandingan Algoritma Rabin-Karp Dan Levenshtein Distance Dalam Menghitung Kemiripan Teks,” *J. Tek. Inform. Unika St. Thomas*, vol. 02, pp. 24–32, 2017.

[39] P. A. R.-K. dan P. P. dengan A. K.-M.-P. Andres, Christopher, and H. Saloko, “Penelaahan Algoritma Rabin-Karp dan Perbandingan Prosesnya dengan Algoritma Knut-Morris-Pratt,” *2006*, no. m, pp. 1–4.

[40] Y. A. Wicaksono, “Analisis Dan Implementasi Algoritma Rabin-Karp Dan Algoritma Stemming Nazief-Adriani Pada Sistem Pendeteksi Plagiat Dokumen,” Bandung, 2012.

[41] T. Mardiana, T. Bharata Adji, and I. Hidayah, “Stemming Influence on Similarity Detection of Abstract Written in Indonesia,” *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 1, p. 219, 2016.

[42] P. Y. Kusmawan, U. L. Yuhana, and D. Purwitasari, “Aplikasi Pendeteksi Penjiplakan pada File Teks dengan Algoritma Winnowing,” pp. 1–11, 2011.

[43] M. Ridho, “Rancang Bangun Aplikasi Pendeteksi Penjiplakan Dokumen Menggunakan Algoritma Biword Winnowing,” PEKANBARU, RIAU, 2013.

[44] A. T. Wibowo, K. W. Sudarmani, and A. moesriami Barmawi, “Comparison Between Fingerprint and Winnowing Algorithm to Detect Plagiarism Fraud on Bahasa Indonesia Documents,” pp. 128–133, 2013.

[45] T. Y. Mahendraputra, “Improvement In Document Similarity Calculation Using Hashing Algorithm And Semantic Analysis On Indonesian Documents,” Universitas Gajah Mada, 2015.

[46] S. Soleman, “Experiments on the Indonesian Plagiarism Detection using Latent Semantic Analysis,” in *2014 2nd International Conference on Information and Communication Technology (ICoICT) Experiments*, 2014, pp. 413–418.

[47] T. Mardiana, “Mesin Pengindikasi Kemiripan Untuk Dokumen Berbahasa Indonesia,” Universitas Gajah Mada, 2015.

[48] U. Taufiq, “Pendekatan Deteksi Plagiarisme Berbasis Kutipan Dan Algoritme Kang Untuk Teks Berbahasa Indonesia,” Yogyakarta, 2015.

[49] T. Mardiana, T. B. Adji, and I. Hidayah, "The Comparation of distance-based similarity measure to detection of plagiarism in Indonesian text," *Commun. Comput. Inf. Sci.*, vol. 516, pp. 155–164, 2015.

[50] J. Kasprzak, M. Brandejs, and M. Kˇripaˇ, "Finding plagiarism by evaluating document similarities," *CEUR Workshop Proc.*, vol. 502, pp. 24–28, 2009.

[51] M. Koppel and J. Schler, "Computational Methods in Authorship Attribution," *Bulg. J. Agric. Sci.*, vol. 60, no. 1, pp. 9–26, 2017.

[52] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," *Am. Assoc. Artif. Intelli- gence (www.aaai.org).*, vol. 12, no. 1, pp. 57–66, 2003.

[53] D. Gupta, K. Vani, and C. K. Singh, "Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection," in *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, 2014.

[54] A. H. Osman, N. Salim, and M. S. Binwahlan, "Plagiarism Detection Using Graph-Based Representation," *J. Comput.*, vol. 2, no. 4, pp. 36–41, 2010.

[55] T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection," *IEEE Trans. Neural Networks*, vol. 20, no. 9, pp. 1385–1402, 2009.