THE IMPLEMENTATION OF ALGORITHM LESK TO SYNONIM RECOGNITION AND RABIN KARP IN DETECTION PLAGIARISM

Moch Nurhalimi ZD[1], Ednawati Rainarli[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail : nurhalimm@yahoo.co.id[1], irene_edna@yahoo.com[2]

ABSTRACT

Plagiarism documents digital is not difficult , enough with using a technique copy-paste-modify .One method used to detect plagiarism is a string of macthing .One algorithm string matching often used is rabin karp according to one journal algorithm rabin karp produce for a great time in detecting string having a pattern of many .Sometimes plagiarism was also done to replace words containing words synonym , to teach look different of the document original .So after the preprocessing process was completed the introduction of synonym and selection said synonym (synonim recognition) .Needed a process to detect in common documents that is the process preprocessing , synonym recognition , parsing k-gram , conversion hashing , matching string , count similiarity .

Based on testing can be concluded that a combination of algorithmic lesk to the process of synonym recognition and rabin rabin karp in the implementation of the in the case of plagiarism (in common in the percentage in common flattened flattened is as much as 85,78 % , than not using a process synonim recognition to create flattened flattened of 77.45 % , although took the process more than do not use synonym recognition

Keyworad : *string matching*, Plagirisme, *Synonim recognition* , algoritma *rabin karp*.

1. INTRODUCTION

Plagiarism or plagiarism means follow or imitative the literature and the work of others who then recognized as karanganya themselves with or without the consent of the author .Cribbing documents digital is not difficult , enough with using a technique copy-paste-modify in some of the and even a whole the contents of of documents was it can be said that the document is the result of duplicate of other documents [1] .Sometimes plagiarism was also done to replace words containing words synonym , to teach look different of the document original .Where only change containing synonym that is in original documents with change his in accordance with existing words di Kamus bahasa indonesia , tanpa pay attention to sentence structure of the original document [1].In this study, there beberapan stages in the process of detecting the similarity of documents from two (2) pieces of the document. Among these processes is the first first stage is the stage of the selection of original documents and document test will count value kesamaanya, afterwards through the stage prepcoessing where the results of this phase will be used for the selection process and determine which word will be selected for completing the follow-plagiarism on changes synonyms using algorithms lesk. Then do the parsing process k-gram of the second array of words (original documents and document test), wherein for determining the value of K filings are experimental, but if the smaller the value of K, the value similiairity will be greater, and the greater the value of K, the value similiairy will the smaller [2]. Then the results of parsing ¬k-gram to be converted to the value of hashing where the value that will be on the skewer using rabin algorithm Karp. And then the calculation process similiairity value of the similarity test document to the original document using the method similiairity dice. The stage pendeteksiap plagiarism of the original document and test documents can be seen in Figure 1 below:

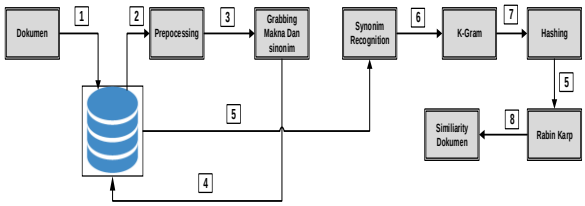


Figure 1. Phase Detection of Plagiarism

1.1 Preprocessing

Documents will be processed, first performed the first stage is the process of preprocessing. The preprocessing phase stages of preprocessing process that will be conducted are splitting the sentence [3], folding case [3], filtering [3], tokenizing [4], stopword removal [5] and stemming [6]. The process is done by using algorithms stemming porter stemming Indonesian [6]. The results of the preprocessing stage of the process will be used to melakuka this process will be used to synonym recognition process by using algorithms lesk.

1.2 Synonim Recognition

Synonym recognition or recognition of synonyms is one approach that is used to assist in the detection of plagiarism. Synonym recognition is a semantic approach to the text document. This approach utilizes the similarity of meaning in the words that are likely to occur [1]. The algorithm used in the process is an algorithm lesk synonym recognition.

Lesk algorithm is an algorithm that is used to remove the ambiguity in the meaning of the word. Lesk algorithm is an algorithm to solve the problem of word sense disambiguation with the dictionary. This algorithm works by comparing the definition of the word definisI.definisi berambigu with its neighbors based on the definition of the word dictionary. Lesk algorithm using WordNet as a dictionary or reference

1.3 Parsing *K-Gram*

K-Gram is a series of terms of length K. Most are used as terms is the word. K-Gram is a method that is applied to the generation of words or characters. K-Gram method is used to pick up the pieces of a k-case characters of a word that continuity is read from the source text to the end of the document [2].

1.4 *Hashing*

Hashing is a way to transform a string into a value that is unique to a certain length (fixed-length) that serves as a marker of the string. Hash function or hash function is a way of creating a fingerprint of various data inputs. Hash function will replace or-kan transposing the data to create a fingerprint, called a hash value [7]. To perform the conversion using equation (1) *below:*

$hash = C1 * a^{(k-1)} + C2 * a^{(k-2)} + C3 * a^{(k-3)} + Ck * a0$

Dan persamaan (2) Berikut

rem = hash / Nilai mod.

Where ci is the ascii character value, k is the k-gram parameter values used and a is a value basis.

1.5 Pencocokan String Dengan *Rabin Karp*

Basically, the Rabin-Karp algorithm will compare the hash values of the input string and substring in the text. If the same, then the comparison will be made once again to the characters. If not the same, then the substring will be shifted to the right. The main key performance of this algorithm adala efficient calculation terhadapa hash value substring when the shift is done.

Due to the algorithm should compare the pattern to the text modulonya same result but different hash values. To avoid unnecessary match, Singh and Kochar (2008) provides a solution to not only compare to the rest of the results, but comparing the results for him as well [8].

$\text{REM}(n1/q) = \text{REM}(n2/q)$ (3)

$\text{QUOTIENT}(n1/q) = \text{QUOTIENT}(n2/q)$ (4)

So successful hit must meet two requirements, namely the remainder of the quotient value and return value must be the same for him. The rest is unsuccessful hit without the need to match again. This means there is no wastage of time to check spurious hit.

1.6 Dice's Similiairity Coeficients

To calculate the similarity of fingerprints obtained the documents used Dice Similarity coeficients by calculating the value of the number of K-Gram used in both documents being tested, while the fingerprint documents obtained from the total value of the K-Gram same. Similarity value can be calculated using the following equation:

$$S = \frac{2C}{A+B} \quad (5)$$

Where S is the value of similarity, and C is the number of K-Gram same on the two texts were compared, while A, B is the number of K-Gram of each of the texts are compared [5].

2. CONTENTS OF RESEARCH

2.1 Data Input

Analysis of the input data is comprised of two documents, the original documents and test documents (documents that have been manipulated) with berektensi file formats .doc and .docx, documents taken from the background of the thesis Informatics Engineering. In the document the test results of his manipulation of data that is changing the word - a word that is a synonym said that the document did not test the same as the original document.

Table 1. Data Input

Dokumen Asli	Dokumen Uji
Plagiarisme merupakan salah satu ***tindakan*** yang ***dilarang***, karena tindakan tersebut termasuk pelanggaran terhadap hasil ***karya*** seseorang yaitu dengan menjiplak atau mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***ijin***. Selain itu, plagiat dapat menurunkan kreativitas	Plagiarisme ialah salah satu ***perbuatan*** yang ***diharamkan***, karena ***perbuatan*** merupakan pelanggaran terhadap hasil ***ciptaan*** seseorang yaitu dengan mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***persetujuan***. Selain itu, plagiat dapat menurunkan
seseorang dalam menciptakan hasil karya. Plagiarisme sering terjadi di berbagai kalangan. Pendeteksian plagiarism dapat dilakukan dengan ***cara*** pencocokan string (*string matching*). Metode ini dapat digunakan untuk menghitung kemiripan	kreativitas seseorang dalam menciptakan hasil ***ciptaan***. Plagiarisme sering terjadi di berbagai kalangan[1]. Pendeteksian ***penjiplakan*** bisa dilakukan dengan menggunakan ***teknik*** pencocokan string (*string matching*).

Dokumen Asli	Dokumen Uji
teks antara satu dokumen dengan dokumen lainnya.	

2.2 Tahap Preprocessing

Preprocessing stage which will be the separation of the sentence, case folding, filtering, tokenizing, stopword removal and stemming. The understanding, examples and simple overview of the preprocessing stage can be seen in Table 2 through Table 7.

1. Separation Sentence

Tabel 2 . Separation Sentence

Sebuah Paragraf	Hasil Pemisahan Kalimat Uji		
	No	**Kalimat**	**Ref**
Plagiarisme ialah salah satu ***perbuatan*** yang ***diharamkan***, karena ***perbuatan*** merupakan pelanggaran terhadap hasil ***ciptaan*** seseorang yaitu dengan mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***persetujuan***. Selain itu, plagiat dapat menurunkan kreativitas seseorang dalam menciptakan hasil ***ciptaan***. Plagiarisme	1	Plagiarisme ialah salah satu ***perbuatan*** yang ***diharamkan***, karena ***perbuatan*** merupakan pelanggaran terhadap hasil ***ciptaan*** seseorang yaitu dengan mengakui hasil ***karya*** orang lain sebagai hasil karya sendiri tanpa adanya ***persetujuan***.	Tidak
Plagiarisme sering terjadi di berbagai kalangan[1]. Pendeteksian ***penjiplakan*** bisa dilakukan dengan menggunakan ***teknik*** pencocokan string (*string matching*).	2	Selain itu, plagiat dapat menurunkan kreativitas seseorang dalam menciptakan hasil ***ciptaan.***	Tidak
	3	Plagiarisme Plagiarisme sering terjadi di berbagai kalangan[1].	Ya
	4	Pendeteksian ***penjiplakan*** bisa dilakukan dengan menggunakan ***teknik*** pencocokan string (*string matching*).	Tidak

2. Case Folding

Tabel 3 . *Casefolding*

Hasil Pemisahan Kalimat Asli		
No	**Kalimat**	**Ref**
1	Plagiarisme merupakan salah satu ***tindakan*** yang ***dilarang***, karena tindakan tersebut termasuk pelanggaran terhadap hasil ***karya*** seseorang yaitu dengan menjiplak atau mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***ijin***	Tidak
2	Selain itu, plagiat dapat menurunkan kreativitas seseorang dalam menciptakan hasil ***karya***.	Tidak
3	Pendeteksian plagiarism dapat dilakukan dengan ***cara*** pencocokan string (*string matching*).	Tidak
4	Metode ini dapat digunakan untuk menghitung kemiripan teks antara satu dokumen dengan dokumen lainnya.	Tidak

3. Filtering

Tabel 3 . *Filtering*

No	Kalimat sebelum *filtering*	Kalimat setelah *filtering*
1	plagiarisme merupakan salah satu ***tindakan*** yang ***dilarang***, karena tindakan tersebut termasuk pelanggaran terhadap hasil ***karya*** seseorang yaitu dengan menjiplak atau mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***ijin***	plagiarisme merupakan salah satu ***tindakan*** yang ***dilarang***, karena tindakan tersebut termasuk pelanggaran terhadap hasil ***karya*** seseorang yaitu dengan menjiplak atau mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ***ijin***
2	selain itu, plagiat dapat menurunkan kreativitas seseorang dalam menciptakan hasil ***karya***.	selain itu, plagiat dapat menurunkan kreativitas seseorang dalam menciptakan hasil ***karya***
3	pendeteksian plagiarism dapat dilakukan dengan ***cara*** pencocokan string (*string matching*).	pendeteksian plagiarism dapat dilakukan dengan ***cara*** pencocokan string string matching
4	metode ini dapat digunakan untuk menghitung kemiripan teks antara satu dokumen dengan dokumen lainnya.	metode ini dapat digunakan untuk menghitung kemiripan teks antara satu dokumen dengan dokumen lainnya

4. Tokenizing

Tabel 5 . *Tokenizing*

No	Kalimat	Proses Tokenizing
1	plagiarisme merupakan salah satu tindakan yang	1. plagoarism 2. merupakan

No	Kalimat	Proses Tokenizing
	dilarang, karena tindakan tersebut termasuk pelanggaran terhadap hasil karya seseorang yaitu dengan menjiplak atau mengakui hasil karya orang lain sebagai hasil karya sendiri tanpa adanya ijin	3. salah 4. Satu 5. tindakan 6. yang 7. dilarang 8. karena 9. tindakan 10. tersebut 11. termasuk 12. pelanggaran 13. terhadap 14. hasil 15. karya 16. seseorang 17. tindakan 18. tersebut 19. termasuk 20. pelanggaran 21. terhadap 22. hasil 23. karya 24. seseorang

5. Stopword Removal

Eliminate words - words contained in the dictionary stopword, such as prepositions and conjunctions.

table 6 . *Stopword Removal*

No	Hasil Stopword
1	1. plagiarism 2. salah 3. satu 4. tindakan 5. dilarang 1. tindakan 2. termasuk 3. pelanggaran 4. terhadap 5. hasil

No	Hasil Stopword
	1. karya 2. menjiplak 3. mengakui 4. hasil 5. karya

Gambar 2. Stopword Removal

6. Stemming

Change the word berimbuhan into basic words.

Table 7 . *Stemming*

No	Array Kata	Hasil *Stemming*
1	6. plagiarism 7. salah 8. satu 9. tindakan 10. dilarang 6. tindakan 7. termasuk 8. pelanggaran 9. terhadap 10. hasil 6. karya 7. menjiplak 8. mengakui 9. hasil 10. karya	1. plagiarism 2. salah 3. satu 4. tindak 5. larang 11. tindakan 12. termasuk 13. langgar 14. terhadap 15. hasil 11. karya 12. jiplak 13. akui 14. hasil 15. karya

2.3 Phase Synonim Recognition

At this stage the introduction stage and choice of words synonyms synonyms to replace the word in the original document and said array array of words in the document so that the test can detect the degree of substitution of the word plagiarism adapaun flowchart synonym stage lesk recognition algorithm can be seen in Figure 2 below :

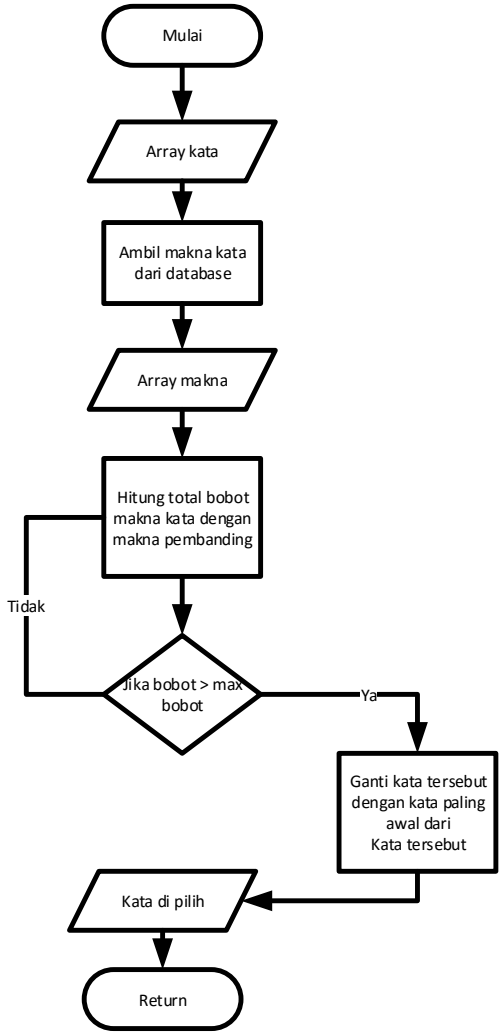


Figure 2 Flowchart Synonim recognition

The results of phase synonym Recognition process can be seen in Figure 3 to the table

Kata Sinonim	Makna Kata	Kata Pembanding	Makna Kata Pembanding
Buat	sesuatu yang diperbuat (dilakukan) untuk menghasilkan sesuatu	Sikap	berbuat tidak peduli atau peduli terhadap apa yang akan terjadi menimpa dirinya dan sebagainya
		Larang	memerintahkan supaya tidak melakukan sesuatu; tidak memperbolehkan berbuat sesuatu
		Tindak	sesuatu yang dilakukan perbuatan yang dilaksanakan untuk mengatasi sesuatu
		Termasuk	sudah masuk ke dalam sesuatu
		Terhadap	kata depan untuk menandai arah
		Hasil	sesuatu yang diadakan oleh seseorang
		Karya	hasil perbuatan buatan ciptaan (terutama hasil karangan) yang bukan tiruan
		Izin	pernyataan mengabulkan yang tidak melarang dan sebagainya
Tindak	sesuatu yang dilakukan perbuatan yang dilaksanakan untuk mengatasi sesuatu	Sikap	berbuat tidak peduli atau peduli terhadap apa yang akan terjadi menimpa dirinya dan sebagainya
		Larang	memerintahkan supaya tidak melakukan sesuatu; tidak memperbolehkan berbuat sesuatu
		Tindak	sesuatu yang dilakukan perbuatan yang dilaksanakan untuk mengatasi sesuatu
		Termasuk	sudah masuk ke dalam sesuatu
		Terhadap	kata depan untuk menandai arah
		Hasil	sesuatu yang diadakan oleh seseorang

Gambar 3 Proses Perhitungan Kesamaan Makna kata

At this stage of the calculation of the weight of the meaning of the word synonyms and word meaning his words will digukana comparison to determine which word will be selected to replace the word synonyms based on the highest weights.

Table 8. Weight calculation

Kata Sinonim	Kata Pembanding	Nilai Bobot
Buat	Sikap	1
	Larang	2
	Tindak	4
	Termasuk	0
	Terhadap	1

2.4 Tahap Parsing *K-gram*

After doing the introduction stage of determining the selection of synonyms and synonyms to address the issue by using the word substitution kamuss synonyms, the next step is parsing kgram, which break down words into pieces where each piece contains a character as much as k.

Table 9 . Parsing K-Gram

Array Kata	Hasil Parsing K-Gram
Plagiarisme adalah satu perbuatan	1. Plag 2. lagi 3. agia 4. giar 5. iari 6. rism 7. isme 8. smes 9. mesa 10. esal 11. sala 12. alah 13. lahs 14. shsa 15. hsat 16. satu 17. atub 18. tubu 19. ubua

Array Kata	Hasil Parsing K-Gram
	20. buat 21. uatl 22. atla

2.5 Phase *Hashing*

After collection - gam collection formed the next step is to change the collection - a collection of the gram to form hash, channeled through the hashing process.

Exsample Array "plagi"

$= (112*10^3) + (108*10^2) + (97*10^1) + (103*10^0)$

$= (112,000) + (10,800) + (970) + (103)$

$= 123,873$

Nilai Mod = 123,873 mod 101

= 47

Reminder = 123,873 / 101

= 1,226.47

As examples of the conversion process all gram to a hash value can be seen in Table 10 below:

Table 10 . *hashing*

Konversi *hashing*			
Array K-Gram	**Nilai Hashing**	**Mod**	**Reminder**
1. plag 2. lagi 3. agia 4. giar 5. iari	1. 123873 2. 118835 3. 108447 4. 114584 5. 115945	1. 249 2. 59 3. 276 4. 252 5. 98	1. 1226.46 2. 1176.58 3. 1073.73 4. 1134.49 5. 1147.97

2.6 String Matching With *Rabin Karp*

After making the changes in the value of k-grams in the form of hash, then the next stage is the process of matching the same hash value from the collection of the hash value test text and set the hash value original text by using an algorithm rabin - Karp, If the hash value of the document being tested same then, the meal will be checked against the value of the remainder of his, her later .Its going to count how many values similiaty hash value of the second set. The flowchart of the string matching process can be seen in Figure 4 below

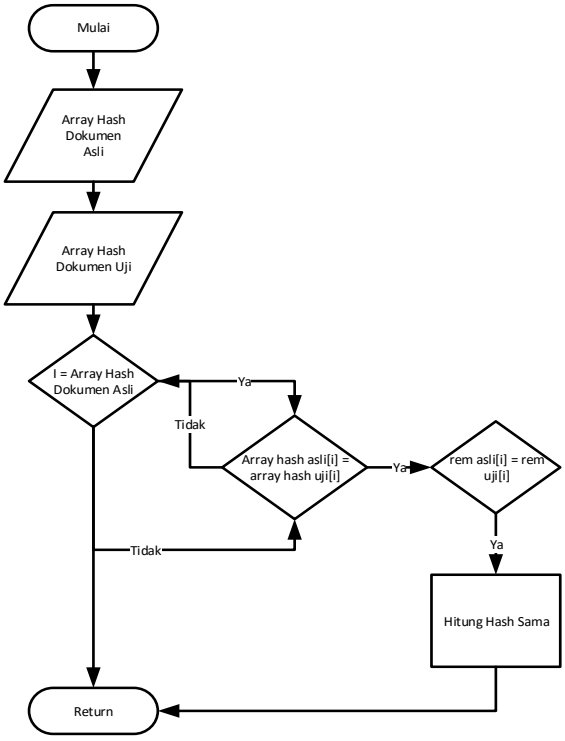


Figure 4 Proses String macthing

Proses Pencocokan String						
Nilai Hashing	Mod	Reminder	Nilai Hashing	Mod	Reminder	Match
1. 123873 2. 118835 3. 108447 4. 114584 5. 115945	1. 249 2. 59 3. 276 4. 252 5. 98	1. 1226.46 2. 1176.58 3. 1073.73 4. 1134.49 5. 1147.97	1. 123873 2. 118835 3. 108447 4. 114584 5. 108874	1. 249 2. 59 3. 276 4. 252 5. 198	1. 1226.46 2. 1176.58 3. 1073.73 4. 1134.49 5. 1246.30	1. Ya 2. Ya 3. Ya 4. Ya 5. No

Gambar 5 Hasil Proses Pencocokan String

Where the results of the string matching that produces the same string is a total of 39 pieces of the original number of 54 pieces of hash and hash test a number of 49 pieces.

2.7 Phase Calculation of Similarity (Dice's Similarity)

After the matching process string, then the similarity value calculation stage. To calculate the similarity of fingerprints obtained the documents used Dice Similarity coeficients by calculating the value of the number of K-Gram used in both documents being tested, while the fingerprint documents obtained from the total value of the K-Gram same.

$$S = \frac{2(39)}{52+47} = \frac{78}{99} = 0.7878$$

Or 78.78 %

3. CLOSING

3.1 Conclusion

Based on the research that has been done, it can be concluded that the results of testing a number of parameter values K-Gram (K = 2, k = 3, k = 4, k = 5) it can be concluded that the combination algorithm lesk to process synonym recognition and rabin rabin Karp in the application in cases of plagiarism (similarity) to produce a percentage of the average - average of 85.78% compared with not using a combination of algorithms lesk untun synonim recognition process which produces only a percentage of 77.45%. although it takes a process that is higher than that do not use synonym recognition.

3.2 Suggestion

Based on the research that has been done, still need to do some further study. The suggestions for further research are:Perlu adanya proses pendeteksian plagirisme untuk mengatasi perubahan kalimat aktif dan pasif.

1. There should be a process to detect plagiarism on the rate of change in the structure of the sentence (pharafrase).Perlu adanya proses untuk menangani, plagiarisme pada perubahan urutan kalimat.
2. There should be a process to replace, word that means the uptake of foreign languages.

BIBLIOGRAPHY

[1] I. a. I. C. S. Dewanto, "DETEKSI PLAGIARISME DOKUMEN TEKS MENGGUNAKAN ALGORITMA RABIN-KARP DENGAN SYNONYM RECOGNITION," *Program Studi Ilmu Komputer, Program Teknologi Informatika dan Ilmu Komputer , Universitas Brawijaya Malang.*.

[2] A. M. Surahman, "PERANCANGAN SISTEM PENENTUAN SIMILARITY," *Program Studi Teknik Informatika.*

[3] W. E. W. d. K. L. Masayu, "Update Summarization Untuk Kumpulan Dokumen Berbahasa Indonesia," *Jurnal Cybermatika*, vol. vol. 1, p. No 2, Desember 2013.

[4] "F. Henry Dan Z. Ery," *Klastering Dokumen Berita dari Web menggunakan Algoritma Single Pass Clustering*, vol. 18, pp. pp. 80-90, 2013.

[5] E. Nugroho, "Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoriam Rabin Karp," *Universitas Brawijaya*, 2011.

[6] A. Ledy, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," *Konferensi Nasional Sistem dan Informatika*, 2009.

[7] H. B. Firdaus, "Deteksi Plagiarisme Dokumen Menggunakan Algoritma Rabin-Karp," *Program Studi Teknik Informatika Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung.*

[8] R. S. C. A. B. Kochar, "RB-Matcher. String Matching Technique," *Rem (Text)*, vol. 234567, pp. no 11, p.3, 2008.

[9] P. Pitria, Analisis Sentimen Pengguna Twitter pada Akun Resmi Samsung Indonesia dengan menggunakan Naive Bayes, Perpustakaan Unikom, 2014.

[10] R. S. C. a. B. Kochar, "RB-Matcher. String Matching Technique," *Rem (Text)*, Vols. 234567, no 11, p.3, 2008.