

Similarity measurement algorithms of writing and image for plagiarism on Facebook's social media

M A Kurniawan¹ and K Surendro²

Institut Teknologi Bandung, Jalan Ganesha No 10, Bandung, Indonesia

¹ aghust.kurniawan@gmail.com, ² surendro@gmail.com

Abstract. Currently, social media is growing very rapidly as a medium of information and interaction. However, these developments commonly not followed by a system that can detect the validity of the information also where the source came from, that resulting a negative potential such as fake news (hoaxes) or plagiarism. As a solution, in this research, will be testing plagiarism of writing and image on social media up by using URL as input test. Methods of measuring similarities that used in this research are Smith-Waterman Algorithm and Latent Semantic Analysis. The research measures the accuracy of both methods with local alignment and term-document approaches on documents and Facebook's social media. The result of the research for Smith-Waterman Algorithm method with local alignment approach is better on document testing, with performance level up to 99.77%. In addition, in the proposed algorithm, image testing performed using Google Image Search can also indicate its relevance to a post. Thus, this algorithm has the opportunity to be develop again in other similarity searches such as authorship identification on a post as well as a hoax analyser.

1. Introduction

Social media is an internet-based media technology that connects every user to participate, share and create content such as blogs, social networks, forums, encyclopedias and virtual worlds. Various kinds of technological devices are currently able to access information on social media with ease due to the availability of an increasingly affordable Internet network. So then, the data spread on the internet more easily accessible and used as a reference. However, sometimes it is become misused by taking the work of others as a property of their own intentionally or not. It is included in the act of plagiarizing without permission or without the inclusion of reference sources commonly known as plagiarism.

Prevention of plagiarism has been widely practiced in previous research. Some tools for identifying plagiarism have been widely developed. Basically comparing only two documents in syntactic resemblance, more precisely looking for similarities based on words, phrases or sentences [1].

Currently, many tools also applied some document similarity detection such as WriteCheck, PlagScan, Turnitin, iThenticate, Viper, etc. In this research, will be submitted an algorithm to test writing and an image on social media as illustration supporting writer. The results of the study expected to provide opportunities for the development of this algorithm with the method of measuring similarity more widely.



2. Plagiarism Detection Methods

2.1. Textual Similarity Approach

The detection of document plagiarism has several approaches in measuring similarity, can be categorized into three types: String-based, Corpus-based, and Knowledge-based [2]. The measurement of textual similarity has an important role in textual relations research and its application in information retrieval, text classification, document clustering, topic detection, topics search, question building, answer questions, essay grading, short answer assessments, translator engines, text summarization and so on.

2.2. Smith-Waterman Algorithm

Smith-Waterman Algorithm uses local alignment techniques to determine similar parts between two sequences of nucleoid acid sequences or protein sequences. Compared to searching through the sequences, this algorithm compares segments with varying lengths that allow and finds optimal alignment. The scheme for obtaining an optimal alignment value can be defined in equation (1).

$$\begin{aligned}
 H(i, 0) &= 0, \quad 0 \leq i \leq m \\
 H(0, j) &= 0, \quad 0 \leq j \leq n \\
 H(i, j) &= \max \left\{ \begin{array}{l} 0 \\ H(i-1, j-1) + s(a, b) \\ \max_{k \geq 1} \{H(i-k, j) + W_k\} \\ \max_{l \geq 1} \{H(i, j-l) + W_l\} \end{array} \right. \begin{array}{l} \text{Match/Mismatch} \\ \text{Deletion} \\ \text{Insertion} \end{array} \right\}, \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (1)
 \end{aligned}$$

Equation (1) will yield an appraisal matrix from the comparison of two series, so that the optimum value obtained which shows the value of similarity between the two. The flow process of this method illustrated by Figure 1.



Figure 1. The Process flow of Smith-Waterman Method

2.3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a statistical model of language learning and representation that uses vector averages in the semantic space to test the similarity between words or text [3]. The use of LSA was first use by Foltz for grading essays in 1995. In addition, several other examples of use as representative models and measuring human verbal concepts [4].

LSA uses Singular Value Decomposition (SVD) to decompose the input matrix which is the matrix of document-term weights into three matrices [5]. The process flow can be illustrate as shown in Figure 2.

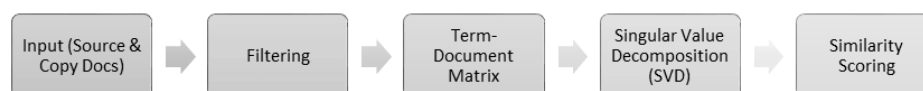


Figure 2. The Process flow of Latent Semantic Analysis

2.4. Performance Measurement

In a documentary measurement system or plagiarism detection, as a classification system determines a pair of documents considered to contain plagiarism. However, the assessment should be support by performance measurement of the method used, to see how accurately it measures the similarity.

Therefore, some performance measurement techniques are need to look at precision, recall and F-Measure values as a measure of an accuracy.

3. Methodology

The design of model architecture in this research use research methodology of Design Science Research Methodology (DSRM) to test case of plagiarism in document and writing in social media. Steps in DSRM include (1) identifying problems and motivations, (2) defining goals for solutions, (3) design, (4) testing, (5) evaluation and (6) communication. These stages are the workflow process conducted on the research and will be combine with the process of filtering and document similarity measurement method.

The design results produce a model that classifies the process into the DSRM step to simplify the testing process. The model shown in Figure 3.

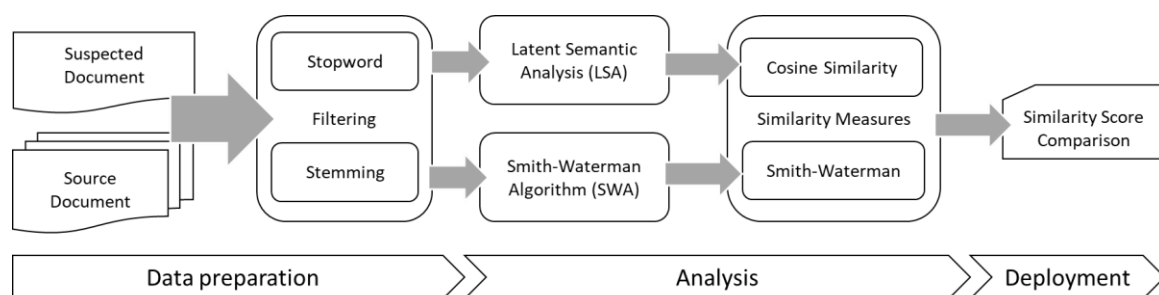


Figure 3. Architecture Model of Plagiarism Testing

In the filtering process, the process of eliminating words that do not have a link of meaning as a whole (Stopword filtering) and eliminate affixes to get the root word (stemming process). Then the measurement of document resemblance done in the analysis phase, to generate the value of similarity.

4. Testing and Implementation

4.1. Testing Similarities on Documents

Tests of similarity in the document process with three scenarios, including testing the same document, testing different documents and examining documents similar to changes in word layout or phrase (paraphrase). In the same document test, both methods show the same results; while for other tests shown in the Table 1.

Table 1. Testing on different documents

No	Copy Doc	Source Doc	Similarity Score	
			LSA	SWA
1	A01	A04	75%	4%
2	A01	A12	81%	2%
3	A01	A14	92%	5%
4	A02	A04	70%	3%
5	A02	A05	91%	6%
6	A02	A11	84%	3%
7	A02	A13	65%	6%
8	A02	A17	87%	4%
9	A03	A11	72%	3%
10	A03	A13	63%	3%

No	Copy Doc	Source Doc	Similarity Score	
			LSA	SWA
11	A03	A16	83%	3%
12	A03	A20	87%	2%
13	A03	A24	82%	2%
14	A04	A09	64%	3%
15	A04	A12	87%	2%
16	A04	A14	74%	2%
17	A05	A08	71%	3%
18	A05	A11	87%	3%
19	A05	A13	71%	6%
20	A05	A17	85%	4%
21	A06	A07	60%	3%
22	A06	A13	67%	3%
23	A06	A21	80%	3%
24	A06	A23	75%	2%
25	A07	A09	62%	3%
26	A07	A10	91%	6%
27	A07	A23	86%	4%
28	A08	A11	78%	3%
29	A08	A13	76%	3%
30	A09	A18	80%	2%
31	A10	A23	75%	4%
32	A10	A28	71%	4%
33	A11	A13	92%	3%
34	A11	A17	74%	2%
35	A11	A21	69%	2%
36	A12	A14	88%	3%
37	A13	A21	65%	2%
38	A15	A22	70%	5%
39	A15	A27	84%	5%
40	A15	A29	84%	6%
41	A16	A20	90%	5%
42	A16	A24	99%	37%
43	A17	A19	73%	2%
44	A17	A21	61%	2%
45	A20	A24	87%	4%
46	A22	A25	77%	3%
47	A22	A27	97%	6%
48	A25	A27	76%	3%

The percentage of similarity that use as a reference is the result of testing on the Latent Semantic Analysis (LSA) method, with an average percentage of similarity 78% because the term frequency in the compared documents has similarities although the topic of the document is different. While the test

results with Smith-Waterman Algorithm (SWA) shows a small alignment value or can be said average percentage of similarity 4%.

Assuming the results of the tests conducted objectively by the authors as a reference level of accuracy, then the results show the average percentage of similarity between the documents close to 0% (no resemblance), can be seen graphically on the Figure 4.

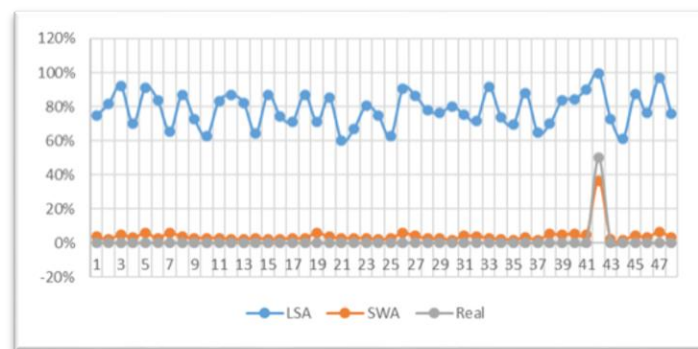


Figure 4. Testing on different documents

In testing similar documents, the test results of both methods are not significantly different, as shown in Figure 5.

	Similarity Score									
	B01		B02		B03		B04		B05	
	LSA	SWA	LSA	SWA	LSA	SWA	LSA	SWA	LSA	SWA
B01	100%	100%	100%	77%	87%	100%	87%	94%	96%	100%
B02			100%	100%	87%	100%	87%	95%	96%	100%
B03					100%	100%	99%	97%	76%	100%
B04							100%	100%	75%	93%
B05									100%	100%

Figure 5. Testing on Similar Documents

From Figure 6 shows that for testing with Smith-Waterman Algorithm, the percentage of similarity is 95.6%. While for Latent Semantic Analysis also shows an overall percentage of 89% average.

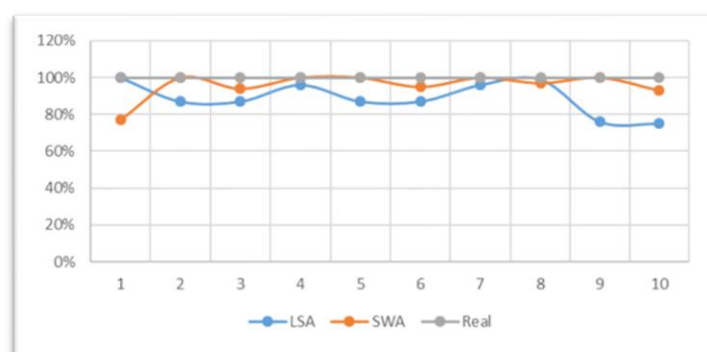


Figure 6. Testing on Similar Documents

4.2. Implementation on Social Media Testing Similarities

Implementation testing of similarity writing and images on Facebook's social media done with some initialization process. The process structured into an algorithm that can extract the content contained in

Facebook's social media. Figure 7 shows an algorithm used to detect plagiarism in Facebook's social media.

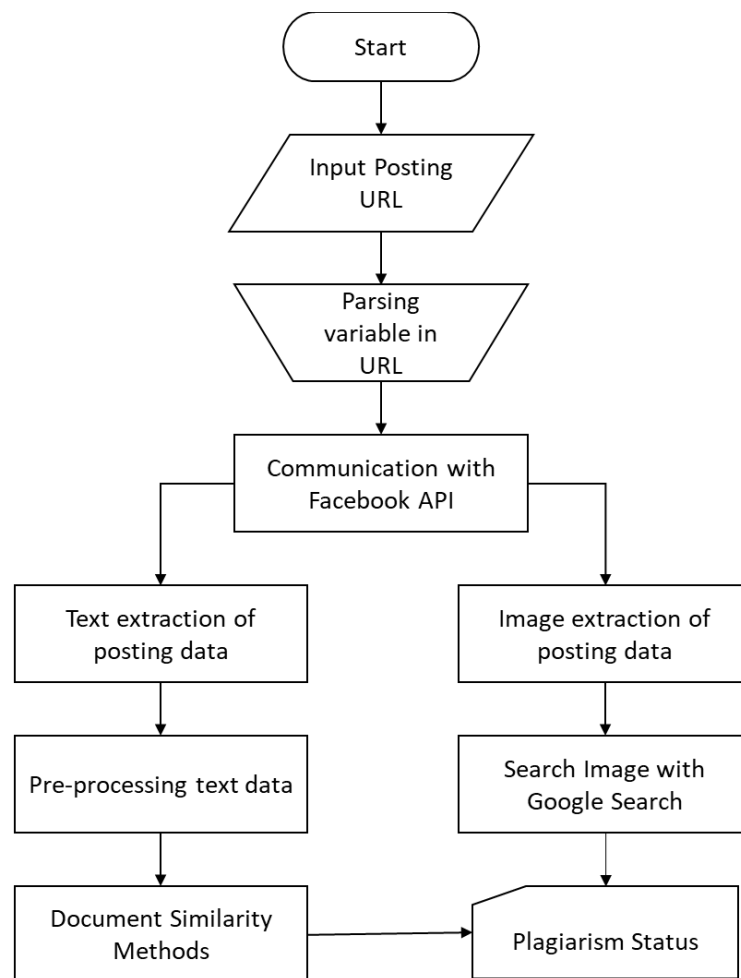


Figure 7. Plagiarism Detection Algorithm on Facebook's social media

Figure 9 is an example of a plagiarism case on Facebook's social media and blogspot. Testing done by initiating the data access of both social media. In this research, the tools used is R, the initiation process, the measurement of similarity and the performance measurement done using the tools. The result of measurement on three cases of plagiarism in social media as shown in Figure 8.

No	Copy Post	Source Post	Similarity Score		
			LSA	SWA	Real
1	F1	F2	100%	67%	60%
2	F3	F4	100%	27%	40%
3	F5	F6	100%	44%	70%

Figure 8. Testing on social media

F3 Agama Facebook Anggaplah engkau teresat di sebuah perkampungan primitif yang sangat jauh dari peradaban. Saat itu, di sana sedang terjadi kehebohan diantara penduduknya. Mereka semua saling berdebat tentang "facebook", sebuah kata di secarik surat botol yang terdampar di pesisir laut. Ada yang menganggapnya hantu, ada yang berpendapat sosok dewa, ada yang ngotot itu surga, ada yang mengira sebuah aliran sesat, ada yang bilang sebuah agama baru, bahkan ada yang tidak percaya bahwa facebook ada. Lalu dirimu yang sebenarnya sudah terbiasa menggunakan facebook mencoba menjelaskannya pada mereka. Tapi tetap saja, mereka masih ngotot dengan pendirian mereka masing-masing. Mereka tak percaya padamu. Penjelasanmu tidak masuk akal. Perdebatan terus berlanjut. Bahkan ada yang sampai bermusuhan setelahnya. Kau akhirnya memilih diam. Kau merasa lebih baik menyaksikan tingkah mereka sambil geleng-geleng kepala. Bagaimanapun, penjelasanmu tidak akan berguna selama mereka tidak melihat facebook dengan mata kepala. Kau hanya bisa berharap semoga listrik dan jaringan internet bisa segera masuk ke kampung tersebut, agar mereka bisa melihat sendiri apa yang sedang mereka ributkan ini. - Afi	F4 Agama Facebook Iadi begini... Anggaplah kau teresat di sebuah perkampungan primitif. Sangat jauh dari peradaban. Saat itu penduduknya sedang ada kehebohan. Sesama mereka saling berdebat tentang "facebook". Ada yang menganggapnya hantu, ada yang berpendapat sosok dewa, ada yang ngotot itu surga, ada yang mengira sebuah aliran sesat, ada yang bilang sebuah agama baru, bahkan ada yang gak percaya facebook ada. Lalu kau yang sebenarnya sudah terbiasa menggunakan facebook mencoba menjelaskannya pada mereka yang sesungguhnya. Tapi tetap saja mereka dengan pendirian mereka masing masing. Mereka tak percaya padamu. Penjelasanmu tidak masuk akal. Perdebatan terus berlanjut. Bahkan ada yang sampai bermusuhan setelahnya. Kau akhirnya memilih diam. Kau merasa lebih baik menyaksikan tingkah mereka sambil senyum-senyum sendiri dan geleng-geleng kepala. Lalu demi keselamatan jiwamu, akhirnya kau memutuskan untuk membiarkan mereka. Kau berharap semoga listrik dan jaringan internet bisa segera masuk ke kampung ini, agar mereka bisa merasakan sendiri facebook itu. Mahala... Kau tau, aku pernah mengalami hal serupa. Dengan topik yang berbeda. Suatu saat aku teresat di sebuah fanspage facebook dengan status yang menyudutkan agama yang aku yakini. Perdebatan yang terjadi di kolom komentar panjang sekali, tembus sampai ratusan. Jelas ini hanya debat kusir, saudara seiman yang mengerti agama mencoba meluruskan pada umat lain yang belum pernah mempelajari dan mengamali sendiri. Mahala... jadi saya diam saja. Tanpa menulis komentar. Karena akan percuma. Dan saya yakin, hal yang sama juga terjadi ketika saya menulis status bertopik agama yang akhirnya mengundang perdebatan diantara sesama saudara seiman, selalu ada "silent reader" dari umat yang disinggung dalam status dan komentar. Tapi mereka lebih memilih diam, sambil senyum-senyum sendiri. Sambil ngurut dada. Sudah saatnya mengucapkan mantra, "Accept and release. Forgive and forget. All is well." The Mind Navigator http://rekoholic.blogspot.co.id/2015/02/agama-facebook.html Tanggal Posting: 24 Februari 2015
--	--

Figure 9. Facebook's status and blogspot's writing

From Figure 8, the Smith-Waterman Algorithm method has a high degree of accuracy and precision when paired with actual results. So the category of plagiarism becomes different, whereas the LSA assumes the whole writing as plagiarized, while SWA considers not all plagiarized. The results of these measurements are not yet be accurate. Therefore, it is necessary to measure performance in the case of plagiarism in social media.

In the image testing on Facebook social media, the image extracted the URL data then performed the same image search using Google Image Search, so that the conclusion about the relevance of the article with the image shown. Figure 10 is one example of testing.

Link to Posting	Picture from Posting	Status
https://www.facebook.com/mochamadridwankamil/photos/a.165770243574805.1073741828.163237587161404/929268827224939/?type=3		Gambar yang dimuat relevan dengan tulisan serta tidak memuat unsur plagiarisme
Result from Image search		

Figure 10. Testing on Image Relevance

4.3. Performance Measurement

In the examination of document similarity, it can be seen the level of precision, recall and F-Measure for both methods in measuring its performance in Figure 11 and Figure 12.

Latent Semantic Analysis			Smith-Waterman Algorithm		
Pre	Rec	F-Me	Pre	Rec	F-Me
23,67%	104,76 %	38,61%	95,23%	104,76%	99,77%

Figure 11. Performance of Methods on Documents

From the results of performance measurement, shows the level of precision and recall method Smith-Waterman Algorithm higher than Latent Semantic Analysis. The average harmonic value between precision and recall of high F-Measure or F-Score indicates a good degree of accuracy. So then, it can be concluded that Smith-Waterman Algorithm method has better accuracy.

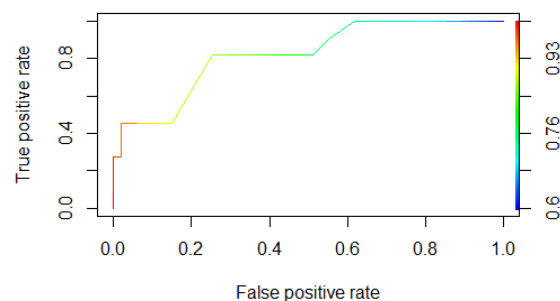


Figure 12. Performance of LSA Method

Performance measurement data are also shown on the graph of true positive rate and false positive rate, where it can be seen that the relevance of document suspected to have similarity/plagiarism is higher in the Smith-Waterman Algorithm method (Figure 13), which is shown with a high true positive rate.

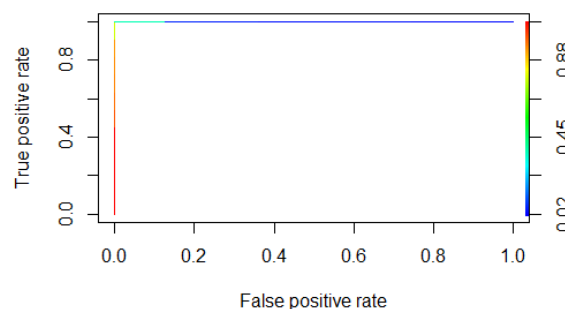


Figure 13. Performance of Smith-Waterman Method

In the performance measurement of similarity test of writing in social media, in Figure 14 show the level of precision and F-Measure Smith-Waterman Algorithm higher, it shows that the accuracy of this method is better.

Latent Semantic Analysis			Smith-Waterman Algorithm		
Pre	Rec	F-Me	Pre	Rec	F-Me
1%	1,76 %	1,27%	2,17%	1,76%	1,93%

Figure 14. Performance of Social Media Plagiarism

5. Results and Suggestions

5.1. Results

Based on the test results, the percentage of high document resemblance does not indicate a good degree of accuracy, a performance measurement process must be performed to determine a good degree of accuracy. After the performance measurement is obtained that the level of precision and high recall indicate a good level of accuracy to the method of measuring similarity of documents.

1. Test results on scientific paper documents and performance measurements, Smith-Waterman Algorithm has a higher accuracy with F-Measure 99.77%.
2. Similar results were obtained on testing similarities of writing data on Facebook social media and performance measurements, Smith-Waterman Algorithm has a higher accuracy with F-Measure value of 1.93% versus 1.27% for LSA.
3. The process of testing the similarity of images based on the extraction of the image is done by the process of execution of the function on R for image search in Google Image Search, because the Google API feature for images is no longer provided by Google. Search function of this image to see the relevance of writing and images on Facebook social media.

5.2. Suggestions

In this study, the authors wanted to show the methods in the measurement of popular document similarities such as Latent Semantic Analysis and Smith-Waterman Algorithm, in addition to having different characteristics, each method has advantages such as the LSA is more efficient when used in testing on large corpus. However, the results obtained show good accuracy on Smith-Waterman Algorithm method by not using corpus.

The suggestion for further research is the development of the Smith-Waterman Algorithm method to perform similarity tests on images in social media as a hoax analyzer. And the implementation of Smith-Waterman Algorithm method in testing the writing pattern with authorship analysis.

References

- [1] A. Sarkar, U. Marjit, and U. Biswas, "A conceptual model to develop an advanced plagiarism checking tool based on semantic matching," *Bus. Inf. ...*, pp. 104–108, 2014.
- [2] W. Gomaa and A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.
- [3] T. K. Landauer and S. T. Dumais, "A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.
- [4] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Process.*, vol. 25, no. 2, pp. 259–284, 1998.
- [5] S. Soleman and A. Purwarianti, "Experiments on the Indonesian plagiarism detection using latent semantic analysis," *2014 2nd Int. Conf. Inf. Commun. Technol.*, no. 2010, pp. 413–418, 2014.