Journal of Applied Data Sciences
Vol. 2, No. 2, May 2021, pp. 45-54

ISSN 2723-6471

Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient

Tri Wahyuningsih [1,*], Henderi [2], Winarno [3]

[1,2,3] Magister Informatics Raharja University, Indonesia
[1] triwahyuningsih@raharja.info*; [2] henderi@raharja.info; [3] winarno1060@yahoo.com
* corresponding author

(Received July 1, 2019 Revised October 21, 2019 Accepted October 29, 2019, Available online October 29, 2019)

Abstract

This study aims to find correlation assessment of Automatic Short Answer Grading (ASAG) by comparing three methods of Cosine Similarity, Jaccard Similarity and Dice Coefficient by providing one reference answer. From the results of computing using Python programming language and data processing using spreadsheets, it was obtained that the Dice Coefficient method had the highest correlation average value of 0.76, followed by Cosine Similarity with an average correlation value of 0.76, and the lowest correlation average value was the Jaccard method with a value of 0.69. The contribution to this study is the use of three methods in one data, whereas the previous research only used 1 method for 1 data or 2 methods for 1 data. So, the value in this study resulted in a more complete comparison and accuracy of data.

Keywords: Text Mining, Automatic Short Answer Grading (ASAG), Cosine Similarity, Jaccard Similarity, Dice's Coefficient

1. Introduction

During the COVID-19 pandemic as it is today, the Indonesian government is taking several policies, one of which is on limiting social interaction. This policy has a significant impact on the world of education. The world of education changed the learning that was originally done on campus but now the learning is done at home so that the learning activities are done online. This encourages educational institutions in Indonesia to start developing e-learning systems in learning activities. E-Learning is one of the learning methods where the learning process, teaching process and even the assessment process are conducted electronically through the internet. By applying e-learning assessment of learning results is done automatically by using automatic grading system. This system has advantages such as being able to score answers quickly and objectively. The assessment model consists of three kinds of multiple choice, right wrong and essay (description) [2]. In college institutions, most lecturers give questions in the form of descriptions. Answers in the form of descriptions are not as easy as answering questions in the form of multiple choice and correct answers are wrong. The answer to the description requires further natural language processing. The answer description is a form of question where the choice of answer is not provided so the student must answer with a sentence. The description answer is the right method to assess the results of the learning activity, because the answer to the description will involve the student's ability to remember and express the ideas they have. The problem in the assessment of the description is about subjectivity, the assessment between one lecturer and another lecturer may be different. Another problem is the possibility of lecturers having errors in assessment such as the answers of the same students but have different scores.

The description assessment system consists of two kinds, namely Automatic Essay Grading (AEG) for essay type and Automatic Short Answer Grading (ASAG) for short answer type. The basic concept of Automatic Essay Grading and Automatic Short Answer Grading is to score on the similarity between student answers and lecturer answers, while the difference between the two lies in the length of the answer. The answer length in Automatic Short Answer Grading ranges from two words to one paragraph [2]. Other researchers limited the number of answers to twenty words so that the results can be more relevant to provide a better correlation [4][5]. The research method often used to score the answers to the description is String-Based Similarity. String-Based Similarity is divided into two characters based and term based. String-Based Similarity is a way to score by calculating the similarity of the character, while term-based calculates similarities based on the terms. ASAG research has been done a lot before, but most datasets used are questions and answers in English form.

The purpose of this study was to provide a score against the similarity of short answers in the type of short answer that uses Indonesian by comparing cosine similarity, Jaccard Similarity and Dice's Coefficient methods.

2. Literature Review

2.1. Text Mining

According to Firdaus [7] text mining is the process of analyzing text to extract useful information for a specific purpose. Text mining is a branch of data mining science. The difference between the two is in the form of data. Data mining has a structured form of data, on the contrary text mining has an unstructured form of data [8]. Research on text mining has been conducted ranging from word matching [9], document compaction [10], plagiarism detection [11], sentiment analysis [12] and automated assessment of essay answers [13][14]. Research on automated assessment using natural language processing techniques was first conducted by page [15]. Then other researchers started to develop it a lot.

2.2. Preprocessing Techniques on Automatic Short Answer Grading (ASAG)

Automatic Short Answer Grading System is an assessment system that is done automatically in the brief by comparing between student answers and lecturer answers. In Table 1. There are five categories of ASAG preprocessing techniques [2].

Table. 1. Automatic Short Answer Grading pre-processing technique

Processing Techniques in Languages	
Lexical	Spelling Correction, Stop word removal
Morphological	Lemmatization, Stemming
Semantic	Anaphora resolution, Named entity tagging, Sense expansion
Surface	Case folding, Number removal, Punctuation removal
Syntactic	Chunking, Parsing, POS Tagging, Sentence Segmentation, Syntactic Template, Tokenization, Word Segmentation

The above categories are presented for four languages namely Chinese, English, Spanish and German. For Indonesian we collected ASAG preprocessing techniques namely Case folding, Tokenization, stop words removal (filtering), and stemming.

2.3. Term-based Similarity Measures Cosine Similarity, Jaccard Similarity and Dice Coefficient

According to Hall and Dowling [16] string-based similarity measurements, it is divided into fourteen algorithms. The fourteen algorithms are seven of them character-based and the other seven term-based. Table 2. Is a similarity measurement based on String.

Table. 2. Measurement of similarity by String

String Based	Algoritma
Character Based	LCS Damerau_ Levenshtein Jaro Jaro Winkler Needleman-Wunsch Smith-Waterman N-Gram
Term Based	Block Distance Cosine Similarity Jaccard Similarity Dice's Coefficient Euclidean Distance Matching Coefficient Overlap Coefficient

For this study focused with three term-based algorithms namely cosine similarity, Jaccard similarity and dice's coefficient.

2.4. Cosine Coefficient Method

Cosine Similarity method is a method used to calculate the similarity between two objects. In general, the calculation of this method is based on vector space similarity measure. This cosine similarity method calculates the similarity between two objects (e.g., D1 and D2) expressed in two vectors using keywords from a document as a size.

2.5. Jaccard Coefficient

Jaccard Coefficient is one of the methods used to calculate similarity between two objects(items). As with cosine distance and matching coefficient, in general the calculation of this method is based on vector space similarity measure.

2.6. Dice's Coefficient

Dice's coefficient is a method for comparing the similarities of two different text samples. Dice coefficient is a semi metric version of Jaccard coefficient. This method maintains accuracy on diverse datasets and gives less weight to datasets containing unrelated features [19].

3. Research Method

This study used the Cosine Similarity, Jaccard Similarity and Dice's Coefficient methods. The programming language used is Python 3.8 to calculate similarity scores and Microsoft Office Excel to calculate correlation and MAE values. Figure 1. The following are the steps taken in conducting research.

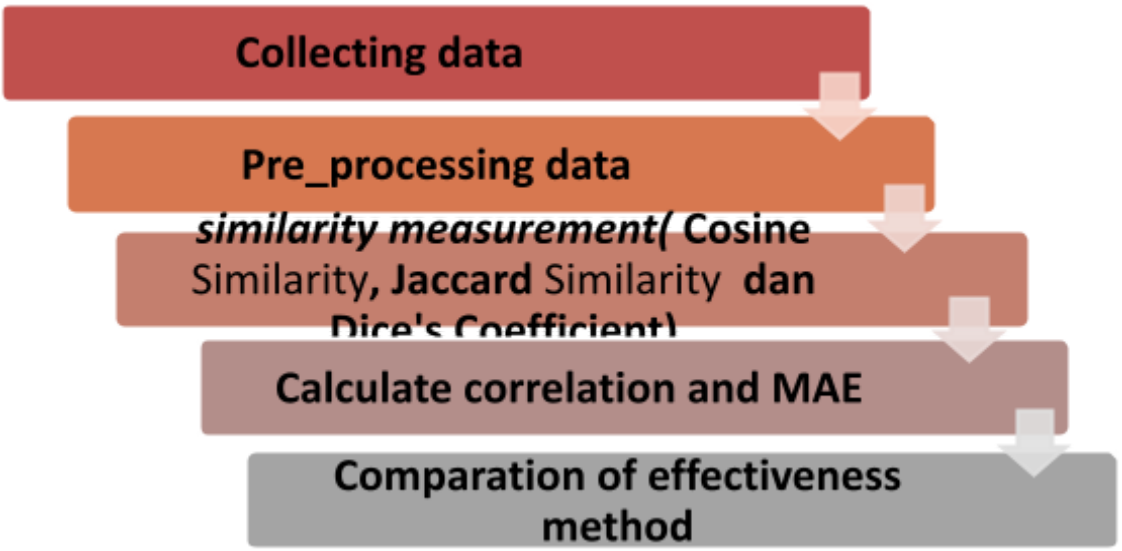

Fig. 1. Research methods and steps

3.1. Collecting data

In this study, the data used were questions and answers in the E-business course quiz at Amikom Purwokerto University which was limited to questions containing definition answers. In this study took four questions answered by thirty-one students each question. Here is an example of questions and answers that can be seen in table 3.

Table 3. Examples of questions and answers

problem	question	Teacher's answer	Student answers
1	What is E-business?	business activities conducted electronically using the internet	business activities conducted electronically or by using the internet
2	What do you know about the B2C sales model?	sales between businesses and consumers directly	The process of buying and selling directly
3	What is E-commerce?	Buying and selling activities of goods or services conducted through the internet	Buying and selling activities conducted through the internet
4	What do you know about B2B?	sales made by the company to other companies, not consumers	inter-company sales

3.2. Pre-processing data

Raw data usually has meaningless parts for text mining, such as stop word. For the data to be processed, the data needs to go through the stage of pre-processing data. Preprocessing data is the process to prepare raw data before the next process is done [20]. The steps in the data processing in this study are:

1. Case folding: case folding is used to convert the entire text to lowercase [21]. This is done to make searching easier, because text documents are not always consistent in the use of letters.

2. Tokenization is the process of dividing text derived from a sentence or paragraph into specific sections [22]. For example, the answer "Direct trade process" generates five tokens namely "Process", "Sell", "Buy", "in", "direct". The separator between tokens is spaces and punctuation.
3. Stop words removal (filtering): In this step will be omitted words that appear frequently but do not contain the meaning [23]. Prefaces and conjunctions are also included in stop words removal.
4. Stemming: The process of converting a word form into a base word [24]

3.3. Cosine Similarity, Jaccard Similarity and Dice Coefficient

The methods used to measure sentence similarity are Word Overlap methods, such as Cosine Similarity, Jaccard Similarity and Dice's Coefficient. This method only counts words that are similar in sentences. The formula of these three methods is shown in table 4.

Table. 4. Formula of Cosine Similarity, Jaccard Similarity and Dice Coefficient algorithms

Similarity method	formula
Cosine Similarity	$\frac{\lvert A \cap B \rvert}{\sqrt{A} . \sqrt{B}}$
Jaccard Similarity	$\frac{\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert - \lvert A \cap B \rvert}$
Dice's Coefficient	$\frac{2 \lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$

3.4. Correlation and MAE

To measure the correlation between teacher answers and student answers, the study used Pearson Correlation on equations (1)

$$r_{xy} = \frac{\sum x.y}{\sqrt{(\sum_x 2)(\sum_y 2)}} \quad \text{.....(1)}$$

In this case:

r_{xy} = Correlation coefficient between variables x and y

X = Deviation from mean value of variable x

Y = Deviation from mean value variable y

$\sum x.y$ = Number of multiplications between x and y values

x^2 = Square of the value x

y^2 = Square of y-value

This study also used MAE (Mean Absolute Error) with the intention to represent the average absolute error between the forecasting result and the actual value [25]. The MAE formula can be seen in the equation (2).

ISSN 2723-6471

$$MAE = \frac{\Sigma[x-y]}{n} \text{ .....(2)}$$

In this case:

MAE = Mean Absolute Error

X = The value of the forecasting result

Y = Actual value

N = Amount of data

The category of success in the automatic scoring system based on the correlation value there are three categories namely Excellent, good and bad. The correlation category is very good the value is r> 0.75, the good category the correlation value is r = 0.40 - 0.75 while the bad category if the correlation value is r <0.4 [26].

4. Discussion

The basic techniques in text mining are tokenization, case folding, stop words removal (filtering) and stemming. In this study removed all punctuation marks and symbols. The teacher and student's answers are input into the token and only take one unique token and then turn it into a vector. After that, change all forms of writing into all lowercase, then stop word removal (filtering) by referring to the research done by Tala [27]. The next stage is stemming by breaking up phrases using the Sastrawi library (https://github.com/sastrawi/sastrawi) based on the Nazief-Adriani Algorithm. In table 5. Describes an example of a preprocessing technique.

Table. 5. Examples of preprocessing techniques

Answer	Business activities conducted electronically using the internet
Case folding	business activities conducted electronically using the internet
Tokenisasi	"Activity", "business", "that", "done", "by", "electronically", "with", "using", "internet"
Stopwords removal	Business activities are carried out electronic internet
Stemming	Enterprising internet business salable electronics

After going through the Pre-processing phase of the data then the next stage measures the similarity of answers using the Cosine Similarity, Jaccard Similarity and Dice's Coefficient methods. In this study, the evaluation metric used was Pearson correlation test. Correlation tests are used to measure the degree of closeness between the value produced by the system and the value provided by the teacher. The assessment is manually done by two teachers, the goal is for the assessment to be done more objectively. The grades given by teachers to students' answers have a range of grades from 0 to 4. Furthermore, the correlation values generated by each method are compared to know the best performance in the assessment of the similarity of the answers. In addition to using Pearson correlation, the study also used Mean Absolute Error (MAE) to measure the error rate between teacher answers and system answers.

Here is an example of calculating the similarity of students' answers to the teacher's answers for the three methods.

A : Enterprising internet business salable electronics

B : Enterprising internet electronic business

ISSN 2723-6471

$A \cap B = 4$

$$Sim_{Cosine} = \frac{4}{\sqrt{5}.\sqrt{4}} = 0,89443$$

$$Sim_{Jaccard} = \frac{4}{5} = 0,80000$$

$$sim_{Dice} = \frac{2\,x\,4}{5+4} = 0,88889$$

The results of the test conducted on four questions with one question each consisted of thirty-one student answers using Cosine Coefficient shown in table 6.

Table. 6. Result of Cosine Coefficient Algorithm

Student Answers	Answer number 1	Answer number 2	Answer number 3	Answer number 4
1	1,00000	0,44721	0,92582	0,70711
2	0,89443	0,67082	1,00000	0,00000
3	0,60000	0,22361	0,61721	1,00000
4	0,47434	1,00000	0,84515	1,00000
5	0,89443	0,91287	0,84515	1,00000
...	...	...	...	...
31	0,44721	0,54772	0,85714	0,57735
r	**0,76163**	**0,76296**	**0,74140**	**0,74329**
MAE	**0,66162**	**0,56689**	**0,55102**	**0,52994**

In table 4 above shows that the highest correlation in cosine similarity method is in answer number 2 with percentage of 0.76296 and lowest in answer number 3 with correlation of 0.74140. As for the highest MAE value shown in answer number 1 and lowest is in answer number 4. The average correlation using the Cosine Similarity method is 0.73804 and the average MAE is 0.57737. Test results with Jaccard Similarity method are shown in table 7.

Table. 7. Result of Jaccard Similarity Algorithm

Student Answers	Answer number 1	Answer number 2	Answer number 3	Answer number 4
1	1,00000	0,40000	0,85714	0,33333
2	0,80000	0,60000	1,00000	0,00000
3	0,60000	0,20000	0,57143	1,00000
4	0,42857	1,00000	0,71429	1,00000
5	0,80000	0,71429	0,71429	1,00000
...	...	...	...	...
31	0,42857	0,42857	0,85714	0,28571
r	**0,70451**	**0,65021**	**0,75863**	**0,67483**
MAE	**0,78816**	**0,79954**	**0,66244**	**0,73925**

In table 5 shows that in this method the highest correlation is found in answer number 3 and the lowest correlation is in answer number 2, while the lowest MAE is shown in answer number 3 and the highest MAE is in answer number

2. The average correlation value for this method is 0.69705 and the mae average value is 0.74735. For test results with Dice Coefficient method shown in table 8.

Table. 8. Result of Dice Coefficient Algorithm

Student Answers	Problem 1	Problem 2	Problem 3	Problem 4
1	1,00000	0,44444	0,92308	0,66667
2	0,88889	0,66667	1,00000	0,00000
3	0,66667	0,22222	0,66667	1,00000
4	0,50000	1,00000	0,83333	1,00000
5	0,88889	0,90909	0,83333	1,00000
...	...	...	...	...
31	0,50000	0,54545	0,85714	0,57143
r	**0,75036**	**0,76010**	**0,70448**	**0,74205**
MAE	**0,65589**	**0,57677**	**0,62200**	**0,53535**

In table 6 shows that in dice's coefficient method has the same correlation value as cosine similarity method which is the highest correlation value found in answer number 2 and lowest in answer three with correlation value of 0.76010 and lowest of 0.70448. For the highest MAE value is in answer number 1 and lowest is in answer number 4. The average correlation value for this method is 0.73925 and the mae average value is 0.59750. For more details the correlation and MAE results comparison for the three methods is illustrated through the bar graph shown in figure 2.

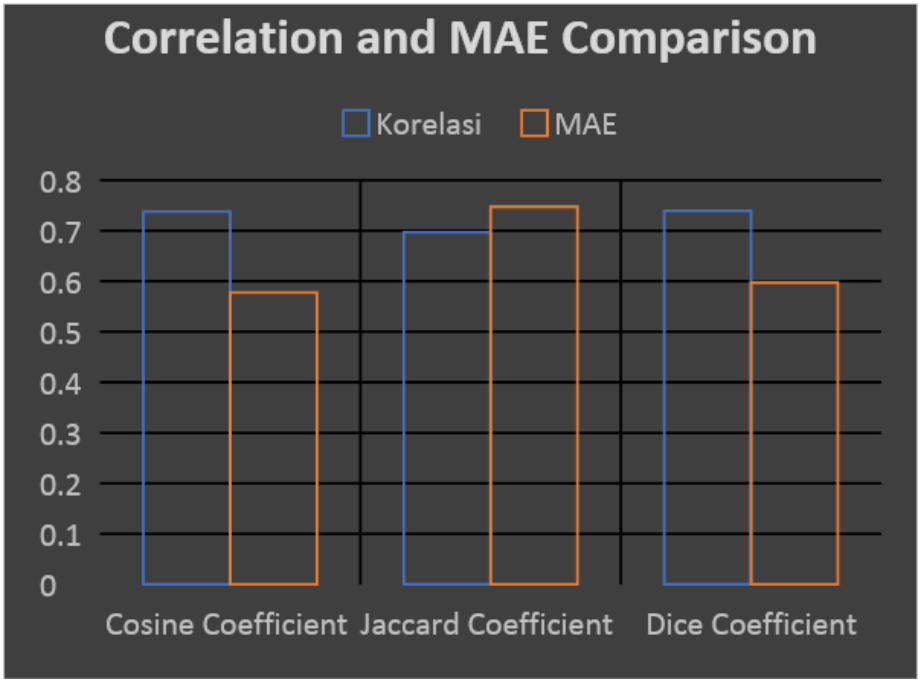

Fig. 2. Comparison of correlation and MAE results

Figure 2 shows that the highest correlation value is found in Dice's Coefficient method with an average value of 0.73925 and the lowest correlation is found in Jaccard similarity method of 0.69705. For the highest MAE is found in Jaccard similarity method of 0.74735 and the lowest is found in cosine similarity method of 0.57737

5. Conclusion

After testing using the Cosine Similarity method, Jaccard Similarity and Dice's Coefficient can be concluded that the highest correlation value is found in Dice's Coefficient method but has a greater MAE when compared to the Cosine Similarity method. The three methods have a correlation between r = 0.40 - 0.75 so that the three methods are said to have a good success rate [26].

References

[1] Putri Ratna, A. A., Budiardjo, B., & Hartanto, D. (2007). SIMPLE : Sistem Penilaian Esai Otomatis Untuk Menilai Ujian Dalam Bahasa Indonesia. Makara, Teknologi, Vol, 11, No.1 , 5-11.

[2] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," Int. J. Artif. Intell. Educ., pp. 60–117, 2015.

[3] V. Salvatore, N. Francesca, & A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading," Journal of Information Technology Education, vol. 2, 2003.

[4] S. Jordan, "Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions," Comput. Educ., vol. 58, no. 2, pp. 818–834, 2012.

[5] S. Jordan, "Short-answer e-assessment questions : five years on," Proc. 15th Int. Comput. Assist. Assess. Conf., 2012.

[6] W. H. Gomaa and A. A. Fahmy, "Short Answer Grading Using String Similarity And Corpus-Based Similarity," Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 11, pp. 115–121, 2012.

[7] Gegick, M., Rotella, P. & Xie, T. 2010. Identifying Security Bug Reports via Text Mining: An Industrial Case Study. IEEE

[8] Imbar, V., Radiant. Adelia, Ayub, M., dan Rehatta, A. 2014. Implementasi Cosine Similarity dan Algoritma Smith Waterman untuk Mendeteksi Kemiripan Teks. Jurnal Informatika Volume 10, Nomor 1.

[9] O. Nurdiana, J. Jumadi, and D. Nursantika, "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia," J. Online Inform., vol. 1, no. 1, p. 59, 2016, doi: 10.15575/join.v1i1.12.

[10] G. Mandar and G. Gunawan, "Peringkasan dokumen berita Bahasa Indonesia menggunakan metode Cross Latent Semantic Analysis," Regist. J. Ilm. Teknol. Sist. Inf., vol. 3, no. 2, p. 94, 2017, doi: 10.26594/register.v3i2.1161.

[11] J. Priambodo, "Pendeteksian Plagiarisme Menggunakan Algoritma Rabin-Karp dengan Metode Rolling Hash," J. Inform. Univ. Pamulang, vol. 3, no. 1, p. 39, 2018, doi: 10.32493/informatika.v3i1.1518.

[12] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," Decis. Support Syst., vol. 48, no. 2, pp. 354–368, 2010, doi: 10.1016/j.dss.2009.09.003.

[13] U. Hasanah and D. A. Mutiara, "Perbandingan metode cosine similarity dan jaccard similarity untuk penilaian otomatis jawaban pendek," Semin. Nas. Sist. Inf. dan Tek. Inform., no. 2019: SENSITIF 2019, pp. 1255–1263, 2019.

[14] S. Roy, S. Dandapat, A. Nagesh, and N. Y., "Wisdom of Students: A Consistent Automatic Short Answer Grading Technique," Proc. 13th Int. Conf. Nat. Lang. Process., pp. 178–187, 2016.

[15] E. B. Page, “Grading Essays by Computer: Progress Report,” Invit. Conf. Test. Probl. 29 October, 1966, vol. 47, no. 5, pp. 87–100, 1966.

[16] P. A. V. Hall and G. R. Dowling, "Approximate string matching, Comput. Surveys", 12:381-402 ,1980.

[17] G. A. Pradnyana dan N. A. Sanjaya, “Cosine Similarity”, Perancangan Dan Implementasi Automated Document Integration Dengan Menggunakan Algoritma Complete Linkage Agglomerative Hierarchical Clustering, vol. 5, (2), pp. 1-10, September 2012.

[18] S. Purwandari, Rancang Bangun Search Engine Tafsir Al-Quran Yang Mampu Memproses Teks Bahasa Indonesia Menggunakan Metode Jaccard Similarity, Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, 2012, pp. 9-27.

[19] Chahal, M. (2016). Information Retrieval using Dice Similarity Coefficient. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, Issue 6, pp.72-75.

[20] Han, J., Kamber, M., & Pei. J. 2012. Data Mining: Concepts and Techniques third edition. Waltham: Elsevier.

[21] Christopher DM, Prabhakar R, Hinrich S. Introduction to Information Retrieval. Introduction to information retrieval. Cambridge University Press. 2008; 1: 496.

[22] Manning, C. D., Raghavan, P., & mSchutze, H. (2009). Introduction of Information Retrieval, Cambridge University Press.

[23] Patel, B., & Shah, D. D. (2013). Significance of stop word elimination in meta search engine. International Conference On Intelligent Systems and Signal Processing (ISSP, 52-55).

[24] G. Carvalho, D. M. de Matos, and V. Rocio, “Document Retrieval for Question Answering : A Quantitative Evaluation of Text Preprocessing,” Proc. ACM first Ph. D. Work. CIKM, pp. 125–130, 2007.

[25] Subagyo, Pangestu, 1986, Forecasting Konsep dan Aplikasi, Yogyakarta, BPFE UGM.

[26] Fleiss J, Levin B, Cho Paik M. Statistical Methods for Rates and Proportions. Third Edit. Technometrics. 2004; 46: 263-264.

[27] Tala FZ. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.Sc. Thesis, Appendix D. Amsterdam. 2003.

[28] Nazief B, Adriani M. Confix Stripping: Approach to Stemming Algorithm in Bahasa Indonesia. Intern Publ Fac Comput Sci Univ Indonesia Depok, Jakarta. 1996;