International Journal of Engineering & Technology, 7 (4.36) (2018) 1386-1390

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Open Problems in Indonesian Automatic Essay Scoring System

Faisal Rahutomo[1]*, Trisna Ari Roshinta[2], Erfan Rohadi[3], Indrazno Siradjuddin[4], Rudy Ariyanto[5], Awan Setiawan[6], Supriatna Adhisuwignjo[7]

[1,2,3,4,5,6,7]State Polytechnic of Malang
**Corresponding author E-mail: faisal@polinema.ac.id*

Abstract

This paper presents open problems in Indonesian Scoring System. The previous study exposes the comparison of several similarity metrics on automated essay scoring in Indonesian. The metrics are Cosine Similarity, Euclidean Distance, and Jaccard. The data being used in the research are about 2,000 texts. This data are obtained from 50 students who answered 40 questions on politics, sports, lifestyle, and technology. The study also evaluates the stemming approach for the system performance. The difference between all methods between using stemming or not is around 4-9%. The results show Jaccard is the best metric both for the system with stemming or not. Jaccard method with stemming has the percentage error lowest than the others. The politic category has the highest average similarity score than lifestyle, sport, and technology. The percentage error of Jaccard with stemming is 52.31%, Cosine Similarity is 59.49%, and Euclidean Distance is 332.90%. In addition, Jaccard without stemming is also the best than the others. The percentage error without stemming of Jaccard is 56.05%, Cosine Similarity is 57.99%, and Euclidean Distance is 339.41%. However, this percentage error is high enough to be used for a functional essay grading system. The percentage errors are relatively high, more than 50%. Therefore this paper explores several ideas of open problems in this issue. The openly available dataset can be used to develop better approaches than the standard similarity metrics. The approaches expose are ranging from feature extraction, similarity metrics, learning algorithm, environment implementation, and performance evaluation.

Keywords: *Indonesian, Natural language processing, Automatic essay scoring system, Open problems.*

1. Introduction

Every learning process requires an evaluation to measure the level of students' understanding. There are many types of evaluations include multiple choice question, short question, and essay question. Some studies have revealed that essay question is better than others if the student's knowledge is evaluated thoroughly [1]. But, the problem arises is time-consuming of the rating process. The teacher should read and evaluate sentence by sentence of student answer.

Nowadays, many information technologies are developed to automate human activities. In the education issue, the developing example is essay grading. Researchers have done research on automated essays scoring (AES) since sixties years last century [2]. There are so many advantages that can be obtained in automated grading rather than in conventional grading. It is reported that teachers in Britain are spending about 30% their time in scoring student's answers and it loses about 30 billion pounds per year [3]. So, there will be many benefits from the application of the automated essay scoring system.

The application of automated essay scoring system has been developed with many different methods being used. However, there is no study indicating which method is better in automated essay scoring, especially in Indonesian. The previous research [4] reveals the average errors of some methods which are commonly used in automated essay scoring in Indonesian. The average errors of each method are calculated with comparing the scores from human raters and scores from the system. The methods are Cosine Similarity, Euclidean Distance and Jaccard. The results show Jaccard is the best approach, but the average error is still high, more than 50%. Therefore this paper exposes several ideas that can be explored further toward this issue. With the benefit of the openly available dataset in http://dx.doi.org/10.17632/6gp8m72s9p.1 [5]. Several evaluations can be done by changing the parameters, such as feature extraction, similarity metric, learning algorithm, environment implementation, and performance evaluation.

This paper presentation is divided into several chapters. Chapter 1 describes the introduction. Then, Chapter 2 exposes the summary of the previous study in English, because Roshinta and Rahutomo report [4] are written in Indonesian. Chapter 3 explores further ideas and open problems toward this issue. Finally, Chapter 4 concludes this paper.

2. Indonesian essay scoring system

Roshinta and Rahutomo [4] propose a web-based automated essay scoring system for Indonesian. The research also develops a dataset for performance evaluation purpose [5]. The study consists of several phases. First, developing the dataset. Inside the dataset are questioned texts with corresponding answer texts. The questions are classified into four categories: lifestyle, politics, sport, and technology. Second, develop the web-based automated essay scoring system. Third, student respondents are asked to answer the questions through web-based application system. Then, the system calculates the score with 3 methods. Fourth, the students' answers are scored manually by 3 lecturer respondents. The final score is defined as the average score of the three respondents then served as the gold standard. Finally, the calculation of the average percentage error between manual scores and the system scores of each method.

Furthermore, this chapter exposes the research summary of Indonesian essay scoring system in English.

2.1. The dataset

The dataset being used in this study is defined in Table 1. The questions are 40 texts which are divided into 4 categories (politics, lifestyle, sport, and technology). Each category has 10 question texts. Roshinta and Rahutomo [4] also provide the answer texts of corresponding questions. An example of Indonesian question text is, "*Jelaskan kegunaan karbohidrat untuk tubuh kita*". The corresponding Indonesian answering text is, "*Fungsi karbohidrat adalah sebagai pemasok energi, dapat memperlancar proses pada pencernaan, memberikan efek kenyang dengan kandungan selulosa-nya dan penyeimbang asam dan basa dalam tubuh*".

The respondents answering the question are around 50 students. The respondents are 2nd grade Information Technology Department student of State Polytechnic of Malang. An example of Indonesian answer text corresponding to above question example is, "*sumber tenaga, pemanis alami, menjaga sistem imun, dan sebagai keseimbangan tubuh*". The total Indonesian answer texts being collected from the respondents are 2,162 texts.

Several problems occur during the answer text collection phase. Therefore several approaches are done toward the issues. The duplicate texts are filtered into a unique text. The problem arises because of duplicate entry into the system by the respondent. Sometimes the student just answers the question carelessly. The problem is investigated further because of the lack of student knowledge toward a general issue in the question text. The other condition is a different amount of corresponding answer text between the questions. The condition happens because the students do not always attend the data collection sessions. Furthermore, the data can be downloaded freely in Mendeley data [5].

Three lecturer respondents give the manual scoring of the students' answers. The score is determined between 0 to 100. Then the final manual score is determined by calculating the average of the three scores.

Table 1: Question dataset [4]

Data	Explanation
Question texts	Total 40 questions, 10 questions of each in category: lifestyle, politics, sport, and technology
Student respondents	2 classes, each around 25 students
Answer texts	2,162

2.1. Text preprocessing

A text of the document can be represented as a vector which each component refers to term [6][7]. The value of this component depends on term existence in a document. Furthermore, not only existence but also it depends on the tern weights which can be obtained from term frequency operation. If the document is represented as a vector, then the mathematic operation can be done. The first process in preprocessing the text is the transformation of text data into numerical data. There are several steps which consist of case folding (convert text into lowercase), tokenizing (explode text into words), stemming (convert words into root word), and stopword (remove words which are not necessary), and term frequency. In the previous study, preprocessing is divided into two ways: with stemming and without stemming.

The tokenizing phase of the text is based on whitespace, with no n-gram consideration. The study uses Nazief and Andriani stemming algorithm [8] in the evaluation system with the stemming process. The algorithm is work for Indonesian. Furthermore, the study uses Tala list [9] in Indonesian stopword phase. Finally, the study calculates the term frequency of the text and generate the term vector of text. The term weights can be obtained from global weighting (by considering the other texts/ document) or local weighting (considering only the text itself). In the study, the term weighting uses local weighting which is expressed by normalized term frequency. The normalized term frequency is a frequency of existence term f_{ij} of term *i* in a document *j* compared with all term in the text [6][7]. Local term weighting of term *i* in document *j* (w_{ij}) can be defined in Equation 1.

$$w_{ij} = \frac{f_{ij}}{\Sigma f_{ij}} \quad (1)$$

Table 2 shows an example of term frequency weight of the term vector in a 5x4 matrix. The rows represent the documents and the columns represent the terms. According to Table 2, the terms are term 1, term 2, term 3 and term 4. The documents are answer key, student1's answer, student2's answer, student3's answer, and student4's answer. Therefore a text is represented as a vector by reading the matrix horizontally.

Table 2: Preprocessing result example [4]

Texts	terms			
	term 1	term 2	term 3	term 4
Correct Answer	0.5	0.2	0	0.3
Student1's answer	0.3	0.4	0.2	0.1
Student2's answer	0.75	0.25	0	0
Student3's answer	1	0	0	0
Student4's answer	0.5	0.5	0	0

2.2. Similarity metric

The study uses three similarity metrics of two vectors: Cosine similarity, Euclidian distance, and Jaccard. The calculation of Cosine Similarity is not derived from the length of the vectors but is derived from the degrees between two vectors [6][7]. The Cosine Similarity can be calculated by Equation 2.

$$Cosine(q,d) = \frac{\sum_{k=1}^{t} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2} . \sqrt{\sum_{k=1}^{t}(w_{dk})^2}} \quad (2)$$

Definition 2.1: w_{ij} = *j*th term weight of document *i*. *q* is a vector of document Q and *d* is a vector of document D.

The value of similarity using Euclidean Distance is done by subtracting the constant (1.42) with a distance of two points of vectors. The constant 1.42 is defined to normalized the result based on the dataset. If the two vectors are completely the same, the result is 1. Contrary, the highest Euclidean Distance of two vectors in the dataset gives score zero. It can be calculated by Equation 3.

$$Euclidean(q,d) = 1.42 - \sqrt{\sum_{k=1}^{t}\left(w_{qk} - w_{dk}\right)^2} \quad (3)$$

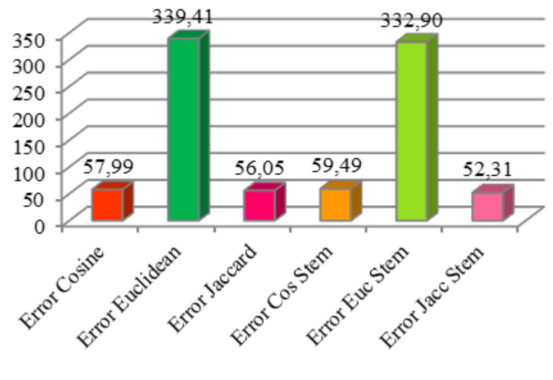
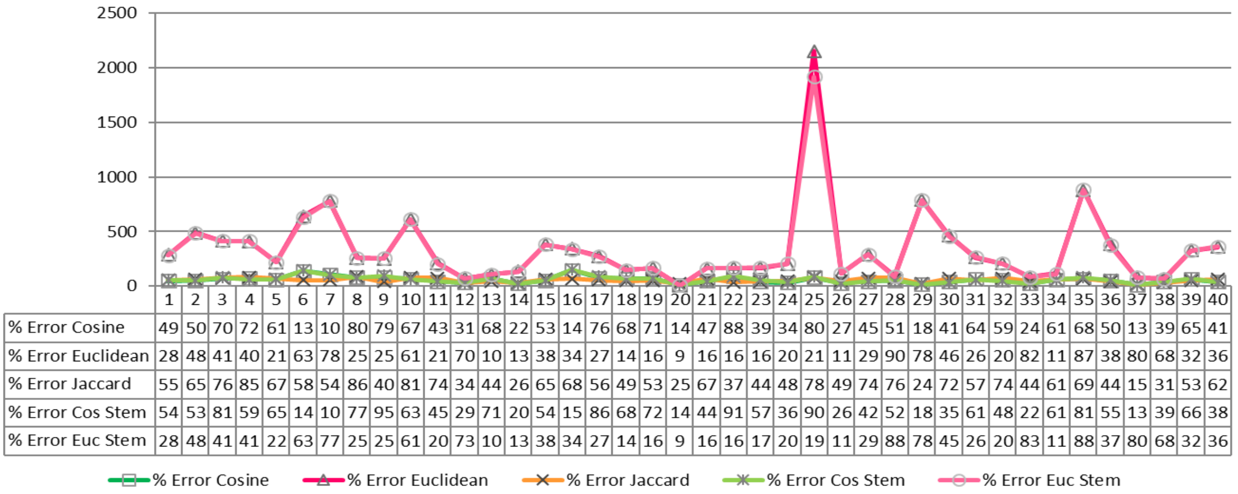


Fig. 1: The percentage error evaluation results [4]



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
% Error Cosine	49	50	70	72	61	13	10	80	79	67	43	31	68	22	53	14	76	68	71	14
% Error Euclidean	28	48	41	40	21	63	78	25	25	61	21	70	10	13	38	34	27	14	16	9
% Error Jaccard	55	65	76	85	67	58	54	86	40	81	74	34	44	26	65	68	56	49	53	25
% Error Cos Stem	54	53	81	59	65	14	10	77	95	63	45	29	71	20	54	15	86	68	72	14
% Error Euc Stem	28	48	41	41	22	63	77	25	25	61	20	73	10	13	38	34	27	14	16	9

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
% Error Cosine	47	88	39	34	80	27	45	51	18	41	64	59	24	61	68	50	13	39	65	41
% Error Euclidean	16	16	16	20	21	11	29	90	78	46	26	20	82	11	87	38	80	68	32	36
% Error Jaccard	67	37	44	48	78	49	74	76	24	72	57	74	44	61	69	44	15	31	53	62
% Error Cos Stem	44	91	57	36	90	26	42	52	18	35	61	48	22	61	81	55	13	39	66	38
% Error Euc Stem	16	16	17	20	19	11	29	88	78	45	26	20	83	11	88	37	80	68	32	36

Fig. 2: The Percentage Error of Each Question [4]

The calculation of Jaccard is described as dividing the number of intersection of terms from two texts by the number of union terms of it [6][7]. Equation 4 describes the formula.

$$Jaccard(q,d) = \frac{q \cap d}{q \cup d} \quad (4)$$

2.3. Percentage error

The calculation of percentage errors of Cosine Similarity, Euclidean Distance and Jaccard can be seen in Figure 1. The gold standard in this experiment is averaged manual scoring by three lecturers as described previously. Figure 1 shows that the Jaccard method with stemming has lowest percentage error, 52.31%. Jaccard without stemming method has an error that is not much different, 56.05%. Jaccard with or without stemming slightly higher than Cosine Similarity. In stemming schema, the difference is around 1.94 %. While in non-stemming schema the difference is around 7.18%. The Euclidean Distance has the highest percentage error. Euclidean Distance without stemming has error around 339.41% and with stemming has error around 332.90%.

Figure 2 shows the percentage errors of each question (1-40). The results clearly show the Euclidean Distance always has the highest error compared with Cosine Similarity and Jaccard. The next subchapters will describe the other slices of analysis to the experiment results.

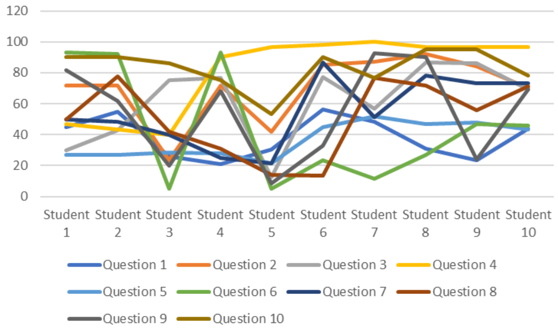


Fig. 3: Student Stability [4]

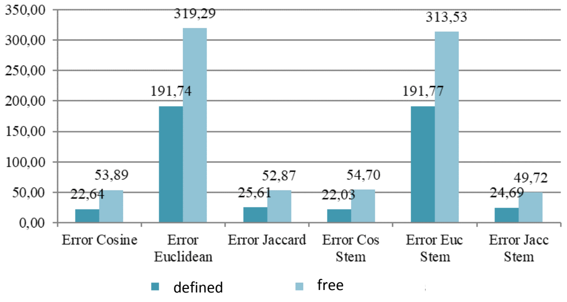


Fig. 4: Percentage Error based on Question Type [4]

2.4. Student stability

Analysis of students' stability shows the score of students. In this analysis, not all the student data will be shown, but only several data taken as samples. Figure 3 shows a graph of students' stability in this study. Figure 3 shows that students occasionally have a high score, but occasionally have a low score. It indicates that students answered the questions base on their ability. There is no pattern showing students always have a high score or low score.

2.5. Percentage error based on question type

Analysis of percentage error base on types of question shows error in question with a free answer and definite answer. In this study, there are 6 questions of the definite answer and 34 questions of the free answer. The numbers of question are not equal, but this study only sees the comparison of each method. Figure 4 shows the percentage error base on types of question. According to Figure 4, the percentage of errors in all methods of definite questions are lower than free questions. Jaccard with stemming method has the lowest error, it is 25.35%. In another hand, Euclidean Distance has the highest error, it is more than 100%.

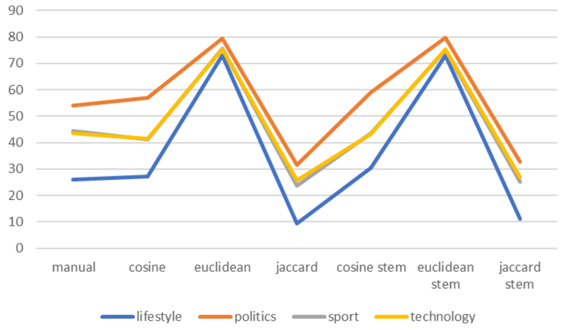


Fig. 5:.Each category average values [4]

2.5. Each category analysis

Analysis of questions' category shows the category with the highest average score. The higher the average score, the better the students' ability in that category. Figure 5 clearly shows that the politics category has the highest average score than lifestyle, sport, and technology. The lifestyle category has the lowest average score. Technology and sports category have slightly differences value.

3. Open problems

This chapter exposes the open problems in this research issue. Due to the lack of experimental results of standard similarity metrics in the previous study, further investigation is needed. The ideas are described as follows.

3.1. Feature extraction

There are many approaches in feature extraction of texts. Several evaluations can be done to the dataset with a different approach. During tokenization of the text, n-gram consideration [10] may be interesting due to the fact, Indonesian lemma is possibly more than one word. Bigram and trigram are considered best for Indonesian. A standardization of terms based on a dictionary [11] is another interesting approach since maybe a respondent did typographical error. Sometimes the respondent is possible to type a correct term of slang or nonstandard term. During standardization of terms by a dictionary, the simplicity of terms by synonym set of thesaurus dictionary [11] and word sense disambiguation technique are possible as well [12]. The removal of unnecessary symbol or extract a correct information with a regular expression is another effort. More advanced preprocessing of terms based on part of speech of terms in a sentence is considerable since the Indonesian part of speech tagger is already available [13].

The previous study uses Nazief and Adriani stemming algorithm for Indonesian [8]. The algorithm obtains the root of Indonesian words. This principle is relatively different than a famous English stemmer algorithm, Porter stemmer. Further research is possible to use Indonesian Porter stemmer algorithm [9] and comparing the results.

If the dimension is an important issue, then stopword removal is an important preprocessing step. The previous study uses Tala list [9]. The other list is available for Indonesian, namely developed by Wibisono [14] and Doyle [15]. Filtering term based on its frequency is possible as well to reduce the dimension. A specific threshold can be defined and several evaluations based on different threshold values are interesting to be investigated further.

In the weighting scheme, global weight such as inverse document frequency (IDF) [16][17] can be evaluated as well. So many weighting schemes are available such as probabilistic retrieval BM25 family [18]. The word2vec vector scheme [19] is interesting as well to be used in this system since matrix of word2vec is a dense matrix, not a sparse matrix like a conventional matrix of term vector of texts.

3.2. Similarity metric

Several distances and similarity metric, different from Cosine, Euclidean, and Jaccard are available as well. For distance scheme, there are the other schemes: Manhattan, Minkowsky, Hamming, Jaro-Winkler, Kendall, Lee, and Levenshtein. For similarity measurement, the other schemes are Dice and Adamic. Further research of that similarity metric performance is interesting as well. A semantic similarity approaches such as latent semantic analysis (LSA) [20] or explicit semantic analysis (ESA) [21] are possible to explore as well. LSA working principle is based on a statistical approach, namely singular value decomposition. Involvement of Indonesian WordNet [22] (if available) with different similarity schemes in a taxonomy such as Wu and Palmer or Lesk is also interesting as well.

3.3. Learning algorithm

Machine learning approaches of classification are seemly working as well in this issue. A quantitative approach like linear regression can be used. With an additional threshold of pass or fail, or marking such as A, B, C, the categorical classification is possible to be evaluated. So many approaches in this approach such as support vector machine (SVM), naive bayes classifier (NBC), decision tree with various variations, KNN, or logistic regression [23][24]. The novel deep learning approach is tempted to be tested in this issue as well [25].

The computational cost in the learning algorithm evaluation can be reduced by dimensional reduction or feature selection. Dimensional reduction of singular value decomposition (SVD) [26][20] or principal component analysis (PCA) [27]are interesting as well.

3.4. Environment implementation

The previous study implements an automatic grading system in a web-based application with PHP CodeIgniter framework. The development of mobile or desktop application is possible as well. Another approach such as front-end and back-end are interesting as well as web service implementation with node js and angular. Several programming languages such as python, java, and VB are another implementation area of exploration.

3.5. Performance evaluation

The previous study only uses average error performance evaluation. Statistical evaluation such as correlation coefficient [28] can be used as well as the other correlation schemes. Standard deviation is interesting as well. Another important evaluation performances are precision, recall, and accuracy [7].

4. Conclusion

This paper has been describing the summary of Roshinta and Faisal study in English. This paper also describes several further research idea toward the issue. Ranging from feature extraction, similarity metric, learning algorithm, environment implementation, and performance evaluation. Hopefully, this paper motivates the other researcher to work in Indonesian automatic essay grading system and improve the learning experience inside the classroom.

Acknowledgment

This study is supported by State Polytechnic of Malang, as a part of innovation research funding with DIPA 042.01.2.401004/2018. The authors would like to thanks to the institution, especially UPT P2M for the full support and funding.

References

[1] M. A. Raihan, R. H. Shamim, C. K. Clement, and H. S. Lock, "A Study on Assessment & Evaluation of Engineering Students' Learning by Essay Test Based on The Cognitive Domain of Bloom's," *Int. J. Adv. Eng. Technol.*, vol. 6, no. 1, pp. 1–11, 2013.

[2] T. Kakkonen and E. Sutinen, "Automatic Assessment of the Content of Essays Based on Course Materials," in *ITRE 2004. 2nd International Conference Information Technology: Research and Education*, 2004, pp. 126–130.

[3] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) System in Indian Context," in *TENCON 2008 - 2008 IEEE Region 10 Conference*, 2008, pp. 1–6.

[4] T. Roshinta and F. Rahutomo, "Analisis Aspek-Aspek Ujian Esai Daring Berbahasa Indonesia," *Pros. Sentrinov (Seminar Nas. Terap.*

Ris. Inov., vol. 2, no. 1, 2016.

[5] F. Rahutomo and T. A. Roshinta, “Indonesian Query Answering Dataset for Online Essay Test System.” .

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed. USA: Addison-Wesley Publishing Company, 2008.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[8] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, “Stemming Indonesian: A Confix-stripping Approach,” vol. 6, no. 4, pp. 1–33, Dec. 2007.

[9] F. Z Tala, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia,” 2003.

[10] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, “Syntactic Clustering of the Web,” *Comput. Networks ISDN Syst.*, vol. 29, no. 8, pp. 1157–1166, 1997.

[11] P. Bahasa, *Kamus Tesaurus Bahasa Indonesia*. Departemen Pendidikan Nasional, 2008.

[12] R. Navigli, “Word Sense Disambiguation: A Survey,” *ACM Comput. Surv.*, vol. 41, no. 2, pp. 1–69, 2009.

[13] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, “Building an Indonesian Rule-Based Part-of-Speech Tagger,” in *2014 International Conference on Asian Language Processing (IALP)*, 2014, pp. 70–73.

[14] Y. Wibisono, “Indonesian Stopword,” 2008. [Online]. Available: https://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/. [Accessed: 01-Aug-2018].

[15] D. Doyle, “Indonesian Stopword.” [Online]. Available: https://www.ranks.nl/stopwords/indonesian. [Accessed: 01-Aug-2018].

[16] G. Salton, A. Wong, and C. S. Yang, “A Vector Space Model for Automatic Indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[17] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Aug. 1988.

[18] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Found. Trends® Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[19] Y. Goldberg and O. Levy, “word2vec Explained: Deriving Mikolov et al.’s Negative-Sampling Word-Embedding Method,” 2014. .

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[21] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis,” in *Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, pp. 1606–1611.

[22] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[23] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006.

[24] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*. Springer Publishing Company, Incorporated, 2006.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[26] V. Klema and A. Laub, “The Singular Value Decomposition: Its Computation and Some Applications,” *IEEE Trans. Automat. Contr.*, vol. 25, no. 2, pp. 164–176, 1980.

[27] I. T. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[28] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson Correlation Coefficient,” in *Noise Reduction in Speech Processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4.