Jurnal Mantik, 5 (2) (2021) 1714-1720

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Using Preprocessing Text Mining With Nazief-Adriani Algorithms Similarity Of Essay Final Exam Semester

Mutiara S. Simanjuntak[1], Joel Panjaitan[2], Syofyan Anwar Syahputra[3]

[1]Information Management, AMIK Universal , Jl. Setia Budi No.90, Medan, 20154, Indonesia
[2,3]Electro Technic, Akademik Teknik Deli Serdang, Jalan Pagar Merbau III Gedung ATDS, Deli Serdang, 20511, Indonesia

E-mail:[1]sarahwaty.mutiara@gmail.com, [2]joel.pandjaitan@gmail.com, [3]anwar.sofyan99@gmail.com

ARTICLE INFO

Article history:
Received: 01 September 2021
Revised: 10 October 2021
Accepted: 01 November 2021

Keywords:
Cosine Similarity, Nazief & Adriani, Database, Keyword

ABSTRACT

The test is one way to measure the level of student ability in participating in learning. One type of exam given to students is the type of essay final exam semester. This research focuses on making automatic grading for essay-type tests using cosine similarity. This method has several stages such as tokenizing, filtering, stemming, analyzing, weighing words in documents with cosine similarity. The stemming process uses the Nazief & Adriani algorithm. The results of this study are concluded that the selection of words that are considered as keywords in the answer key greatly affects the results of the assessment of the system. This is evidenced by testing applying the cosine law of 89.5%. However, there are several types of questions that are significantly different because there are unique characters in the database and answer keys that do not contain keywords that match the correct answer.

1. Introduction

At this time the world is feeling the impact of the Coronavirus Disease (Covid 19) pandemic. Indonesia is one of the countries that are severely affected, especially in the field of education, causing schools and universities to be unable to conduct the learning process face-to-face. Learning is diverted by applying online methods or online learning using media such as google classroom, zoom, WhatsApp, and other methods [1].

The application of online learning methods is also applied to the implementation of midterm exams and final exams. Exam questions given to students can be essays or multiple choices. Multiple choice questions are filled by choosing answers from those provided. Unlike the essay problem that requires students to give the answers, they have to follow the student's understanding. The answers students produce aren't solely right or wrong but there is also the possibility of approaching right. As an application, if the perfect answer is 100 then the wrong answer is given a value of 0 and the answer is close to given the value of 40 etc.

Arifin Noor had conducted research by using the Cosine Equation to students' essay exam difficulties. The method is not used at the Preparatory Text Mining Step, namely at the stemming stage, therefore the process of printing the root word is useless and lacks criterion [2]. The stemming process has an impact on the accuracy of information collection. Stemming is accomplished by eliminating the word's affixes.

The selection of a technique or algorithm for a case must also be exact because it is dependent on the objectives and the accuracy of the results[3]. The authors employed the Prepossessing Text mining step in this i nvestigation, employing the Nazief and Adriani algorithms at the Stemming stage for Indonesian terms. Many algorithms, including the Nazief and Andriani algorithms, have been devised to carry out the Indonesian stemming procedure.

Therefore, to facilitate the process of correcting and assessment, a system that applies algorithms to be able to calculate the similarity of student answers to the answer key that has been provided by lecturers. The

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

authors applied the Cosine Similarity method to analyze students' answers to produce similarities. It was then combined with the Nazief & Adriani algorithm for the process of stemming against words.

1.1 Information retrievals System

Information retrievals System is one of a clump of computer science related to the retrieval of information in the collection of documents both content and context that must be found to realize the desire for users of information [4]. Information that can be obtained from the Information Retrievals System can be text, picture, audio, and video that is useful for searching for information and maintaining information [5].

1.2 Stemming

A process contained in an IR (Information retrieval) system is a stemming process. This stemming process is tasked with transforming the words contained in a document into a root word by applying a certain rule [5]. Stemming is also one of the stages used for booster performance (improving performance) information retrieval in Indonesian text intended to eliminate suffixes, prefixes, and prefixes, certainly different from English text where the stemming process is used to eliminate suffixes [6].

1.3 Nazief -Adriani Algorithm

The Nazief-Adriani algorithm was first developed by Bobby Nazief and Mirna Adriani. Nazief and Adriani's stemming algorithm was developed based on morphological rules Indonesian grouped by prefix, suffix and confixes referred to as conjunctions [7]. The basic word dictionary is used for the Nazief & Adriani Algorithm and is supported by recording, such as the preparation of words that undergo the process of stemming excess. Grouping additions into several categories according to morphological rules Indonesian are as follows [8]:

a. Inflection suffixes are a group of suffixes whose basic word does not change. For example, the word "makan" given the suffrage "-lah" would be "makanlah." This group can be divided into two:
   - *Particle* (P) such as, “pun” ,“tah” ,“-kah” and “-lah”
   - *Possessive pronoun* (PP) such as “-ku”, “-nya” and “-mu”.

b. *Derivation suffixes* (DS) is original indonesian language which added directly to root word such us suffiexes “-kan”, “-an” and “-i”,.

c. *Derivation prefixes* (DP) is an prefixes which added before the beginning (prefixes) or after the end (suffixes). In category such us:
   - “be-“, “te”, “pe-“ and “me” which is morphology prefixes.
   - “ke-“, “se-“ and “di-“ or which is not morphology prefixes.

   The affixed words in Indonesian based on classification affixed can be formulated as: [9]:

$$\left[\mathrm{DP} + \left[\mathrm{DP} + [\mathrm{DP} +]\right]\right] root\ word\ [[+\mathrm{DS}][+\mathrm{PP}]] \qquad (1)$$

   dsc:
   DP: *Derivation prefixes*
   DS: *Derivation suffixes*
   PP: *Possessive pronoun*

The use formulated Nazief & Adriani algorithm is [9]:

a. The combination affixed “se-kan”, “be-i”, “ke-kan” ,“ke-i”, “me-an”, “te-an” and “se-i”.
b. Shouldn’t use affixed repeadly.
c. When a word only consist of one or two font can not be process.
d. The prefixes which added to change original root word or prefixes that before has been given such “me-“ change to “men-“, “mem“, meng-“ and “meny-“. Therefore a rule is needed to address the morphology.

Algoritma Nazief & Adriani who made by Bobby Nazief and Mirna Adriani has processing stages which are described in the by the formula: [9]:

$$Prefiks\ 1 + Prefiks\ 2 + Kata\ dasar + Sufiks\ 3 + Sufiks\ 2 + Sufiks\ 1 \qquad (2)$$

1) First, look for the word to be imported into the root dictionary. If found, it's assumed that the word is the root word. Then the algorithm stop.

Accredited “Rank 4”(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2) Inflection Suffixes ("-lah", "-kah", "-ku", "-mu", or "-nya") deleted. If particles ("-lah", "-kah","-tah" or "-pun") then the steps can be repeated to delete then Possessive Pronouns ("-ku", "-mu", or "-nya"), if there.
3) Deletion of Derivation Suffixes ("-i", "-an" or "-kan"). If the word can be found in a basic dictionary, then the algorithm ceases. If not proceed to c1
   a. If the word "-an" is removed and the last letter of the word is "-k", then "-k" would delete. If the word gets find in the dictionary algorithm is cease. If not found do the next step c2.
   b. The Suffixes is removed ("-i", "-an" or "-kan") will be returned, next to step d
4) Delete Derivation Prefix. if in step c have suffixes deleted then proceed to step d1, if not proceed go to step d2.
   a. Check the combination table prefix - suffix not allowed. If founded then algorithm escase, if isn't
   b. Go to step d2.
5) For i = 1 to 3, define the type of prefix then deleted prefix. if the root word not find doing step e, if has been alogarithm escase. Noted : If the first prefix same with second prefix algorithm stop.
6) Do *Recording*. If all the steps have been completed then it is also not success full then the initial word can be assumed as the root word. Process completed.

1.4 Method *Cosine Similarity*

Cosine Similarity is a same measure which use for retrieval information and measure point of view between vector document D_a (titik (*ax*, *bx*)) dan D_b (titik (*ay* , *by*)). Each vector is represented in every word in the document (text) which is compared to triangel, thus the law cosine can be applied to state that [10]:

$$Similarity = \cos(\theta) = \frac{A.B}{||A|.||B||} \quad (3)$$

Simply *Cosine Similarity* used to compare the level of similarity of documents with the concept of cosine degrees where the result are limited between 0 and 1. The document is declared not similar if the result 0. The document is declared similar if *cosine similarity* 1.

$$\cos(\theta) = \frac{\sum_{i=0}^{n} A.B}{\sqrt{\sum_{i=1}^{n}(A_i)^2}.\sqrt{\sum_{i=1}^{n}(B_i)^2}.} \quad (4)$$

Desc :

A = Student Answer
B = Lecture Answer
A_i = integrity of word i in block A_i
B_i = integrity of word i in block B_i
i = number of words in sentence
n = number of vector

2. Method

The research methodology use in this study is a research method. Qualitative methods are those that aim to understand the phenomena experienced by the research subjects as a whole in the form of words and language in a natural context. The framework of this research is shown in Figure 1.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

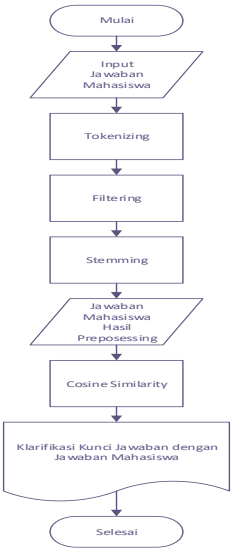

Fig 1. The Research Framework

The stages of research out in this study are:

2.1 Data collection

Collecting the question data and answer key course Data Mining of AMIK Tunas Bangsa that will be tested and collect concepts or supporting theories in this research.

2.2 Essay exam form

Define the right keywords from the answer key as a reference for exam assessment, checking the answer from students by making keywords as a reference for the correct answer then calculating a measure of each question and adding the product of the maximum scores from the question the students that will be finish score.

2.3 Essay exam Architecture

The method used for the answer key with an answer from students as well as improving system function when an error occurs.

2.4 Algorithm implementation

Stages of answers between student answers and answer keys using method *Cosine Similarity* accompanied by the result of data processing. Making an automatic assessment system has several: *tokenizing*, *filtering*, *stemming*, *analyzing*, word bombing in documents with cosine similarity. This process is the process of changing words into basic words. The *stemming* process in the sample test uses the nazief dan adriani algorithm.

3. Result And Analysis

The trial process using 60 samples of students with three different questions then the result of the assessment of the answers is compared with results of student answers using the nazief & adriani algorithm calculation process and combining cosine similarity method. The following are the results of calculation using the nazief & adriani algorithm and *cosine similarity* method, are:

1) *Case Folding* and *Tokenizing*

At the *case folding* stage, changes will be made to all lowercase letters. Can be seen in table 1

TABEL 1
CASE FOLDING

Sentence	Result of *Case Folding*
A (student answer)	IF outlook overcast THEN yes IF outlook rain AND wind false THEN yes OR wind true THEN no IF outlook sunny AND humidity <77.50 THEN no OR humidity <=(is less than or equal to) 77.500 THEN yes.
B (answer key)	IF outlook = overcast THEN yes IF outlook = rain AND wind =false THEN yes OR wind=true THEN no IF outlook = sunny AND humidity =>77.500 THEN no OR humidity <=77.500 THEN yes

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Then proceed with the *tokenizing* stage by breaking the sentence into several words. Tokenizing stages can be seen in table 2:

TABEL 2
TOKENIZING

Sentence	Result of Tokenizing
A (student answer)	IF, outlook, overcast, THEN, yes, IF, outlook, rain, AND, wind, false, THEN, yes, OR, wind, true, THEN, no, IF outlook, sunny, AND, humidity, THEN, no, OR, humidity, is less than or equal to, THEN, yes.
B (answer key)	IF, outlook, overcast, THEN, yes, jika, outlook, rain, AND, wind, false, THEN, yes OR, wind, true, THEN, no, IF, outlook, sunny, humidity, THEN, no, OR, humidity, THEN, yes

2) Filtering

At the *filtering* stage,the process of deleting words that are considered to have no effect on the core of the sentence is carried out. Data beheading is done by deleting the word "di-, ke-, se-". Stages of *filtering* results can be seen in table 3 :

TABEL 3
FILTERING

Sentence	Result of Filtering
A (student answer)	IF, outlook, overcast, THEN, yes, IF, outlook, rain, AND, wind, false, THEN, yes, OR, wind, true, THEN, no, IF, outlook, sunny, AND, humidity, THEN, no, OR, humidity, is less than or equal to, THEN, yes.
B (key answer)	IF, outlook, overcast, THEN, yes, IF, outlook, rain, AND, wind, false, THEN, yes OR, wind, true, THEN, no, IF, outlook, sunny, humidity, THEN, no, OR, humidity, THEN, yes

3) Stemming

At the stemming stage, the word variable is split into basic words. At this stage of stemming using the Nazief & Adriani algorithm process, it can be seen in table 4:

TABLE 4
STEMMING ALGORITMA NAZIEF & ADRIANI

Sentence	Result of Stemming
A (student answer)	IF, outlook, overcast, THEN, yes, IF, outlook, rain, AND, wind, false, THEN, yes, OR, wind, true, THEN, no, IF, outlook, sunny, AND, humidity, THEN, no, OR, humidity, is less than or equal to, THEN, yes.
B (answer key)	IF, outlook, overcast, THEN, yes, IF, outlook, rain, AND, wind, false, THEN, yes OR, wind, true, THEN, no, IF, outlook, sunny, humidity, THEN, no, OR, humidity, THEN, yes

4) Analyzing

At the analyzing stage using the *cosine similarity* method to accive the value of sentences A and B. Stages of analysis obtained from the number of words from Nazief & Adriani algorithm which can be seen from table 5 :

TABEL 5
THE RESULT STEMMING NAZIEF & ADRIANI ALGORITHM

No	Word	A	B
1	if	3	3
2	outlook	3	3
3	overcast	1	1
4	then	5	5
5	yes	3	3
6	rain	1	1
7	and	2	1
8	wind	2	2
9	false	1	1
10	or	3	2
11	true	1	1
12	no	2	2

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

13	sunny	1	1
14	humidity	2	2
15	is	1	0
16	less	1	0
17	than	1	0
18	equal	1	0

After finding the result of the Nazief & Adriani *stemming* algorithm , calculation where carried out using equation 3, namely :

$$Similarity = \frac{(3x3)+(3x3)+(1x1)+(5x5)+(3x3)+(1x1)+(2x1)+(2x2)+(1x1)}{\sqrt{(3+3+1+5+3+1+2+2+1+3+1+2+1+2+1+1+1+1)}\ X}$$

$$\frac{(3x2)+(1x1)+(2x2)+(1x1)+(2x2)+(1x0)+(1x0)+(1x0)+(1x0)}{\sqrt{(3+3+1+5+3+1+1+2+1+2+1+2+1+2+0+0+0+0}}$$

$$Similarity = \frac{77}{9{,}274\ X\ 8{,}602}$$

$$Similarity = \frac{77}{79{,}755}$$

$$Similarity = 0{,}965$$

After performing the process of calculating the *cosine similarity,* the result is 0,965 with a percentage of similarity of answers of 96,5 %.

5) Value calculation process.

The results of the assessment were obtained from one student who obtained the results of the presentation of similarities of 96,5%, 82%, 90% with a total of 3. The following is the calculation of the value:

$$Value = \frac{96{,}5 + 82 + 90}{3} = 89{,}5$$

From the values ,it can be concluded that the average level of results for the calculation of the value is above 80%,meaning that the answer keys and results of the students answers that have been compared have a significant word similarity with total value of yang signifikan 89,5%.

4. Conclusion

The conclusion by applying the *Cosine Simalirity* algorithm and Nazief & Adriani to the Essay Exam Assesment concludes that the choice of words that are considered as keywords in the answer key greatly effect the assessment results of the tests carried out get a match accuracy value of 89,5%.

5. References

[1] W. Darmalaksana, R. Y. A. Hambali, A. Masrur, and Muhlas, "Analisis Pembelajaran Online Masa WFH Pandemic Covid-19 sebagai Tantangan Pemimpin Digital Abad 21," *UIN Sunan Gunung Djati Bandung*, vol. 1, no. 1, pp. 1–12, 2020.

[2] Arifin Noor, "'Automatic Essay Assessment Application Using the Cosine Similarity Method,'" *Shaft Tech*, vol. 7, no. 2, p. http://ejurnal.poliban.ac.id/index.php/porosteknik, 2015.

[3] R. Rosnelly, D. Hartama, M. Sadikin, C. Lubis, M. Simanjuntak, and S. Kosasi, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," *Turkish J. Comput. Math. Educ.*, vol. 12, pp. 1415–1422, Apr. 2021.

[4] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, 2017, doi: 10.30646/sinus.v15i2.305.

[5] A. Prasidhatama and K. M. Suryaningrum, "Perbandingan Algoritma Nazief & Adriani Dengan Algoritma Idris Untuk Pencarian Kata Dasar," *J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, pp. 1–4, 2018, doi: 10.26905/jtmi.v4i1.1773.

[6] H. T. Nugroho, "Pengaruh Algoritma Stemming Nazief-Adriani Terhadap Kinerja Algoritma Winnowing Untuk Mendeteksi Plagiarisme Bahasa Indonesia," *J. Ultim. Comput.*, vol. 9, no. 1, pp. 36–40, 2017, doi: 10.31937/sk.v9i1.572.

[7] M. W. Sardjono, M. Cahyanti, M. Mujahidin, and R. Arianty, "Pendeteksi Kesamaan Kata untuk Judul Penulisan Berbahasa Indonesia Menggunakan Algoritma Stemming Nazief-Adriani," *Sebatik 2621-069X*, pp. 138–146, 2018.

[8] A. Bastian, H. Sujadi, and P. A. Sukmana, "Rancang Bangun Aplikasi Penilaian Ujian Essay Dengan Menggunakan Algoritma Nazief & Andriani Dan Metode Cosine Similarity," *infotech J.*, vol. 4, no. 2, pp. 62–68, 2018.

[9] P. N. Banjarmasin, Y. Anistyasari, E. Hariadi, J. T. Informatika, and U. N. Surabaya, "ALGORITMA BARU PEMBENTUKAN KATA DASAR PADA PROSES STEMMING BAHASA INDONESIA," *Pros. SNRT (Seminar Nas. Ris. Ter.*, vol. 5662, no. November, pp. 70–76, 2019.

[10] M. AGUS SALIM, "Pengembangan Aplikasi Penilaian Ujian Essay Berbasis Online Menggunakan Algoritma Nazief dan Adriani dengan Metode Cosine Similarity," *It-Edu*, vol. 2, no. 01, 2017.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.