Jurnal Mantik
Volume 3 Number 4, February 2020, pp. 268-277 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

# Expert System of Text Mining to Analyze Student Interaction in FTKI UNAS Online Lectures

Novialdy[1], Ucuk Darussalam[2], Ben Rahman[3]

Sistem Informasi
Fakultas Teknologi Komunikasi Dan Informatika , Universitas Nasional
Jl. Sawo Manila, Rt.14/03, Ps. Minggu, Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12520

Email: diboy.Aldi@Gmail.com[1], ucuk.darussalam@gmail.com[2], ben.rahman@gmail.com[3]

ARTICLEINFO

*Article history:*
*Received: 24/01/2020*
*Revised:28/ 01/2020*
*Accepted: 01/02/2020*

***Keywords**:*
*Text Mining, Analisi,*
*Tf-Idf Cosine Similarity.*

**ABSTRACT**

*Interactivity is the student activities or relationships between students with anything related to the course, his example by attending a class attendance responds directly at the appointed hour. What is meant here is the liveliness activities interactivity of students to answer the questions that exist in the online lecture forum, whether he is copying and pasting answers with other students or not. The number of student answers archives make web manager can not judge whether the student was copying and pasting the answer to her or not. Therefore we need a system that can provide a connection weights to archives word answers and calculates similarities between students answer word. To analyze the data from these students, writers manipulate the data using text mining. This research using tf-idf method cosine similarity, the parameters used are folding case, tokenizing, stopwords and stemming. The results of this study are that have a common answer to answer other students not considered effective in the online lectures, corresponding similarity weight percentage of statutes that have been determined.*

## 1. Introduction

Educational institutions in Indonesia continue to improve the poor quality of education, especially college students. One university programs in order to improve the quality of education using online lectures. National education serves to develop the ability and character development and civilization of the nation's dignity in the context of the intellectual life of the nation, aims to develop students' potentials to become a man of faith and fear of God Almighty, morality noble, healthy, knowledgeable, skilled, creative, independent and become citizens of a democratic and responsible (MONE, 2003, p.4). Not many other universities that use these online college system, so a lot of some of the features that did not exist, his example of how active kah Students with online lectures that have been applied.

Student interactivity can be analyzed using several factors aimed at improving the quality, such as the attitude factor in answering questions and activeness. Of these factors can be summed up into a conclusion that later can be seen that these factors affect the student's academic value.

Online lectures are intended in this study is an avenue of learning as well as the usual learning but this learning system through the Web University connected with the internet.

I want analysis research here is related parameters in response to one student activeness learning forum, the forum provided in the system amounted to 3 forum. To analyze the data from these students, writers manipulate the data using text mining. Then at the stage of implementation of the data using the TF-IDF cosine similarity. By using text mining writer able to process and analyze large amounts of data, in whole or in part analyzing unstructured text. And supported by the TF-IDF method Cosine similarity that is able to analyze the active students and students who are not active, with a good degree of accuracy.

Objective of this final task is done to Know the weight of each answer student resemblances to assess whether the answer is copypaste degan other students.

This study refers to previous research that has been done by Budi Santosa, Dwi Smaradahana and with the title "Application of Text Mining for Data Clustering Perform Tweet Shopee Indonesia". In the journal Shopee Indonesian parties can determine the type of content that many do retweeted tweets by followers Shopee Indonesia, so it can use the tweet content types as a means to carry out advertising to users of Twitter [1].

Fauzi Bayu True, Purwono Hendradi, and Bambang Pujiarto entitled "Detection of plagiarism scientific works with the use of bibliography on similar themes search using the cosine similarity". In the journal has a problem with the system that there has been no detection system plagiarism computerized so the researchers aim to design and build applications detection of plagiarism with the use of bibliography using cosine similarity rated for use in detecting the similarity of the text da identify any element of plagiarism between documents. [ 2].

This research was conducted by Ni Luh Ratniasih, Sudarma Made and Nyoman Gunantara entitled "Application of text mining in spam filtering for applications chat". In these journals have problems with spamming action on one of the communication facilities are found on the internet chat feature so that users of these features are not comfortable with their spamming. This research will be to design an a chat application that is able to filter out spam using text mining and engineering challenge response filtering. [3].

This research was conducted by Ria Melita, Victor Amrizal, Hendra Bayu Suseno, and Taslimun Dirjam entitled "Application of term frequency inverse document frequency (tf-idf) and cosine similarity the information retrieval system to find out Sharh hadith web based "on the journal wants to do a retrieval system information to know Sharh hadith, so I wanted to do a retrieval system return information that a web-based using the cosine similarity can be done to search for documents relevant to the we want.[4].

This research was conducted by Rizki Tri Wahyu, Dhidik Prastiyanto, and Eko Supraptono entitled "Implementation of algorithms cosine similarity and weighting tf-idf on the system of classifying documents thesis" problems in such journals is unable Unclassified well with many archive documents the thesis that the accumulated result back search process becomes difficult. Therefore we need a system that can classify documents automatically, using tf-idf method which is a way to give weight to a word against the document, and cosine similarity method for calculating the similarity between a document with other documents. [5].

This research was conducted by Rito Putriwana Pratama, Muhammad Faisal, and Ajib Hanani, entitled "Detection of Plagiarism journal articles using the cosine similarity" to the problem of the journal is the nature of plagiarism is the act of taking an idea or take a research person without citing sources, thus requiring methods which is used to find the weight of similarity between the articles is a method that is based on the cosine similarity of vectors that have similar number of words in the two articles were compared. [6].

Entitled "Implementation of Algorithms Cosine Similarity and TF-IDF Weighting the Final Document Classification system". The problems that exist in many of its journals proficiency level is resulting document archive retrieval process becomes difficult. Then it takes the system to automatically classify documents into a different folder on the database to make it easier manage it. Worn cosine similarity method to compare the similarity advance of title documents with keywords first. [7].

This study refers to previous research that has been done by Rhevitta Widyaning Palupi, Yuita Arum Sari, and the son of Pandu Adikara entitled "Prediction rating new novel is based on a synopsis using the genre-based collaborative filtering and text similarity" in the journal tells of electoral rating on novel novel that is making the reader feel confused to know the quality of the novel. This study uses collaborative filtering-based genre as predictive calculation rating and text similarity to determine the value of the similarity between documents with one another. [8].

The study also refers to the international journal History conducted by Maedeh Afzali and Suresh Kumar, entitled "Comparative Analysis of Various Similarity Measures for Finding Similarity of Two Documents" in these journals have a problem in the measurement in selecting the document so that researchers make the analysis of various sizes in common to evaluate the performance measures this similarity in determining the similarity of two documents. The aim of the journal is to measure the similarity between two documents or documents and request. [9].

The study also refers to the journal History, entitled "Automatic essay scoring system using n-gram and cosine similarity for gamification based e-learning" these journals have a problem with weaknesses in the assessment system is still manual so that it can take a very long time, because the researchers will make the system automatically assessment essay using the cosine similarity and n-grams. [10].

This study refers to the journal History entitled "Using explicit semantic similarity for an improved web explorer with ontology and tf-idf" the journal is to look for extracting and select the hyperlink best and take a more accurate search for the search query is entered. [11] ,

This study refers to the previous journal, entitled "The Analysis of the Proximity Between Subjects Based on Primary Contents Using Cosine Similarity on Lectiv" The journal tells analyzing two subjects by using the cosine similarity and lectiv. The results of this analysis judgment is quite high with 90.91% accuracy in spite of the low recall value so as to facilitate the analysis process between lecturers associated with the subjects in the same do aim. [12].

This study refers to the earlier international journal, entitled "A code classification method based on TF-IDF" The journal has a problem in the code in the document so that the author tried to use the pre-grouping grouping system to extract the relevant features of the document. Then the files are compared with the same name. [13].

This study refers to the previous national journal, entitled "Application of text mining classification system of spam emails using Naive Bayes" The journals have an issue against the spread of information that is not as good as viruses and advertising a company or promoting a specific business products. So users fret to such behavior, then made a research to develop a text mining application that is capable of classifying email. In the email classification probability value is calculated based on the appearance of the word contained in the email data. The results have been made capable of generating accuracy of 89.6%. [14].

The last one of this study refers to the journal earlier is called "text mining and sentiment analysis of twitter in the LGBT movement" of this journal have a problem chirp on twitter associated pros and cons of the LGBT movement, the next step is to analyze text mining worldclouds and continued analysis of sentiment in tweet. So this research could be the referral data from a larger research step. [15].

## 2. Research methods

### 2.1. text Mining

Text mining is a technique that can be used to perform classification, text mining which is a variation of data mining are trying to find interesting patterns from the collection of large amounts of textual data. In the application of text mining, there are several steps that need to be done include:

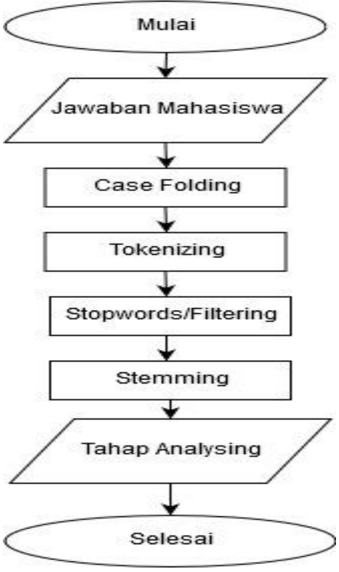


**Fig** 1. The system flowchart.

a) Case Folding
A phase change from uppercase letters to lowercase.
b) tokenizing

Tokenizing the original description decomposition process in the form of a sentence into words.

c) Stopwords / Filtering

Step take the important words from the process token. Could use the algorithm stop list or word list. Filtering can also be defined as the process of taking the word - an important word from the process token or deletion stopwords. Stopwords is a vocabulary that is not a characteristic (unique word) of a document.

d) stemming

Stemming is the stage to find the root word of the results of filtering. Stemming is the process of mapping and decomposition of various forms (variants) of a word to form the basic word (stem).

e) Analyzing stage

Analyzing the stage of determining how far the connectivity between a word or term to a document or sentence by calculating the value / weight of connectedness. Algorithm TF / IDF is used in the process of calculating the weight (W) terminology words. This algorithm is used to calculate the weight of each word most commonly used in information retrieval. This method is also well-known efficient, easy and have accurate results.

### 2.2. The method used:

a. TF-IDF method

TF-IDF is a method for calculating the weight of each word most commonly used in information retrieval. This method is also well-known efficient, easy and have accurate results. In the TF-IDF algorithm used formula to calculate the weight (W) of each document for the keywords with the formula is:

WDT = tfdt * IDFT

Where :
WDT = weight of documents all of the words to the d-t
Tfdt = number of search terms in a document.
IDFT = inversed document frequency (log (n / df))
N = total document
DF = a lot of documents that contain the search terms

b. cosine similarity

Cosine similarity is a measure of similarity between two vectors in a dimensional space obtained from the cosine angle of the multiplication of two vectors being compared because the cosine of 0 is 1, and less than 1 to the value of another angle, then the value of the similarity of vectors is said to be similar when the value of the cosine similarity is 1. The formula is as follows:

$$Cos(): \frac{\sum(Qdf.Qd1)}{\sqrt{\sum Q_{df}2}\sqrt{\sum Qd1^2}} \text{ ............................................. ............................................. ....... (1)}$$

Calculate the similarity vector (document) query Q with every document that there are similarities between documents can use the cosine similarity.

**The first step :**

Calculate the scalar multiplication results between Q and other documents. The result is a multiplication of each document with Q summed (corresponding numerator in the formula above)

**Step Two:**

Calculate the length of each document, including Q. How squared weighting each term in each document, add up the value of the square and the last akarkan.

## 3. Results and Discussion

### A. Use Case Diagram

Use case diagram is a diagram illustrating how a user interacts with the system. Based on observations in Moodle LMS-use case diagram researchers modeled on the system, namely:

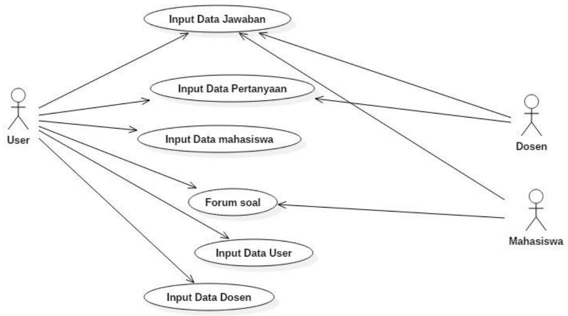
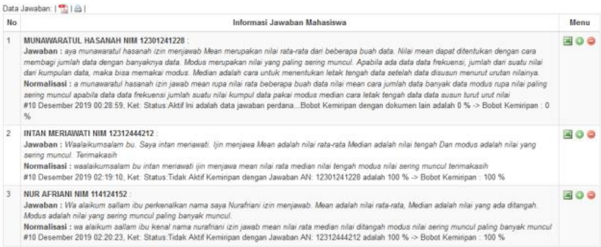


**Fig** 2. Use Case Diagram

## B. input Data

Inputted data is a student response to a question that has been provided by the lecturer in the form of text answers. The following student response data page views:



Data Jawaban: | |

| No | Informasi Jawaban Mahasiswa | Menu |
|---|---|---|
| 1 | **MUNAWARATUL HASANAH NIM 12301241228** :<br>**Jawaban :** *aya munawaratul hasanah izin menjawab Mean merupakan nilai rata-rata dari beberapa buah data. Nilai mean dapat ditentukan dengan cara membagi jumlah data dengan banyaknya data. Modus merupakan nilai yang paling sering muncul. Apabila ada data data frekuensi, jumlah dari suatu nilai dari kumpulan data, maka bisa memakai modus. Median adalah cara untuk menentukan letak tengah data setelah data disusun menurut urutan nilainya.*<br>**Normalisasi :** *a munawaratul hasanah izin jawab mean rupa nilai rata beberapa buah data nilai mean cara jumlah data banyak data modus rupa nilai paling sering muncul apabila data data frekuensi jumlah suatu nilai kumpul data pakai modus median cara letak tengah data data susun turut urut nilai*<br>#10 Desember 2019 00:28:59, Ket: Status:Aktif Ini adalah data jawaban perdana...Bobot Kemiripan dengan dokumen lain adalah 0 % -> Bobot Kemiripan : 0 % | |
| 2 | **INTAN MERIAWATI NIM 12312444212** :<br>**Jawaban :** *Waalaikumsalam bu. Saya intan meriawati. Ijin menjawa Mean adalah nilai rata-rata Median adalah nilai tengah Dan modus adalah nilai yang sering muncul. Terimakasih*<br>**Normalisasi :** *waalaikumsalam bu intan meriawati ijin menjawa mean nilai rata median nilai tengah modus nilai sering muncul terimakasih*<br>#10 Desember 2019 02:19:10, Ket: Status:Tidak Aktif Kemiripan dengan Jawaban AN: 12301241228 adalah 100 % -> Bobot Kemiripan : 100 % | |
| 3 | **NUR AFRIANI NIM 114124152** :<br>**Jawaban :** *Wa alaikum sallam ibu perkenalkan nama saya Nurafriani izin menjawab. Mean adalah nilai rata-rata, Median adalah nilai yang ada ditangah. Modus adalah nilai yang sering muncul paling banyak muncul.*<br>**Normalisasi :** *wa alaikum sallam ibu kenal nama nurafriani izin jawab mean nilai rata median nilai ditangah modus nilai sering muncul paling banyak muncul*<br>#10 Desember 2019 02:20:23, Ket: Status:Tidak Aktif Kemiripan dengan Jawaban AN: 12312444212 adalah 100 % -> Bobot Kemiripan : 100 % | |

**Fig** 3. The data page answers

In figure 3 is a data set answer to answer student who had responded to questions already.

To perform the weighting value of words and find the value of similarity at each student answers, examples of which I took Response answers are tested answer number 3 and then answer No. 3 compared to the previous answer before. to go through the process of folding case, tokenizing, stopwords and stemming. Following the results of the parameters used parameters:

## C. Phase Case Folding

Stage process of folding the above case to change the letters, from uppercase into lowercase letters.

**Table 1.**
Case Folding

| Before In Case Folding | | After In Case Folding |
|---|---|---|
| Alaikum wa sallam mother permission introduce my name Nurafriani answered. The mean is the average value, median is the value that is ditangah. The mode is the value that appears most frequently arise. | | alaikum wa sallam mother permission introduce my name nurafriani answered. the mean is the average value, the median is the value that is ditangah. mode is the value that appears most frequently arise. |

## D. phase tokenizing

Tokenizing the original description peguraian process in the form of a sentence into words.

**Table 2.**
tokenizing

| Before In tokenizing | | After in tokenizing |
|---|---|---|
| Alaikum wa sallam mother permission introduce my name Nurafriani answered. The mean is the average value, median is the value that is ditangah. The mode is the value that appears most frequently arise. | | wa<br>alaikum<br>sallam<br>mother<br>Let me introduce you<br>name<br>I<br>nurafriani<br>permission<br>answer |

| Before In tokenizing | | After in tokenizing |
|---|---|---|
| | | mean |
| | | is |
| | | score |
| | | flat |
| | | flat |
| | | median |
| | | is |
| | | score |
| | | that |
| | | there is |
| | | ditangah |
| | | modus |
| | | is |
| | | score |
| | | that |
| | | often |
| | | appear |
| | | most |
| | | Lots |
| | | appear. |

## E. Phase stopwords / Filtering

Stage took the important words from the process token.

**Table 3.**
Stopwords / Filtering

| Before In stopwords | | After in stopwords |
|---|---|---|
| wa | | wa |
| alaikum | | alaikum |
| sallam | | sallam |
| mother | | mother |
| Let me introduce you | | know |
| name | | name |
| I | | nurafriani |
| nurafriani | | permission |
| permission | | answer |
| answer | | mean |
| mean | | score |
| is | | flat |
| score | | median |
| flat | | score |
| flat | | middle |
| median | | modus |
| is | | score |
| score | | often |
| that | | appear |
| there is | | most |
| ditangah | | Lots |
| modus | | appear |
| is | | |
| score | | |
| that | | |
| often | | |
| appear | | |
| most | | |
| Lots | | |
| appear. | | |

## F. Phase Stemming

Stemming is the stage to find the root word of the results of filtering, or removing the additive.

**Table 4.**
stemming

| Prior Stemming | | After in Stemming |
|---|---|---|
| wa | | wa |
| alaikum | | alaikum |
| sallam | | sallam |

| | |
|---|---|
| mother | mother |
| know | know |
| name | name |
| nurafriani | nurafriani |
| permission | permission |
| answer | answer |
| mean | mean |
| score | score |
| flat | flat |
| median | median |
| score | score |
| middle | middle |
| modus | modus |
| score | score |
| often | often |
| appear | appear |
| most | most |
| Lots | Lots |
| appear | appear |

## G. Weighting Kata (Tf-Idf)

Before calculating the similarity (cosine similarity) between each answer student. Tf-Idf following calculation:

**Table 5.**
TF / IDF

| WORD | TF | | | | IDF | WDT = TF.IDF | | |
|---|---|---|---|---|---|---|---|---|
| | Q | D1 | D2 | DF | Log (N / DF) | QDF | QD1 | QD2 |
| wa | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| wwalaikum | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| sallam | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| mother | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| know | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| name | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| nurafriani | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| permission | 1 | 1 | 0 | 2 | log (3/2) = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 0 x 0.17609125905568 = 0 |
| answer | 1 | 1 | 0 | 2 | log (3/2) = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 0 x 0.17609125905568 = 0 |
| mean | 1 | 2 | 1 | 3 | log (3/3) = 0 | = 1 x 0 = 0 | = 2 x 0 = 0 | = 1 x 0 = 0 |
| score | 3 | 5 | 3 | 3 | log (3/3) = 0 | = 3 x 0 = 0 | = 5 x 0 = 0 | = 3 x 0 = 0 |
| flat | 1 | 1 | 1 | 3 | log (3/3) = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 |
| median | 1 | 1 | 1 | 3 | log (3/3) = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 |
| ditangah | 1 | 0 | 0 | 1 | log (3/1) = 0.47712125471966 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 0 x 0.47712125471966 = 0 |
| modus | 1 | 2 | 1 | 3 | log (3/3) = 0 | = 1 x 0 = 0 | = 2 x 0 = 0 | = 1 x 0 = 0 |
| often | 1 | 1 | 1 | 3 | log (3/3) = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 |
| appear | 2 | 1 | 1 | 3 | log (3/3) = 0 | = 2 x 0 = 0 | = 1 x 0 = 0 | = 1 x 0 = 0 |
| most | 1 | 1 | 0 | 2 | log (3/2) = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 0 x 0.17609125905568 = 0 |
| Lots | 1 | 1 | 0 | 2 | log (3/2) = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 1 x 0.17609125905568 = 0.17609125905568 | = 0 x 0.17609125905568 = 0 |
| a | 0 | 1 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = | = 1 x 0.47712125471966 = | = 0 x 0.4771212547196 |

| WORD | TF | | | | IDF | WDT = TF.IDF | | |
|---|---|---|---|---|---|---|---|---|
| | Q | D1 | D2 | DF | Log (N / DF) | QDF | QD1 | QD2 |
| | | | | | | 0 | 0.47712125471966 | 6 = 0 |
| munawar atul | 0 | 1 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.4771212547196 6 = 0 |
| hasanah | 0 | 1 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.4771212547196 6 = 0 |
| appearan ce | 0 | 2 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 2 x 0.47712125471966 = 0.95424250943932 | = 0 x 0.4771212547196 6 = 0 |
| some | 0 | 1 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.4771212547196 6 = 0 |
| fruit | 0 | 1 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 1 x 0.47712125471966 = 0.47712125471966 | = 0 x 0.4771212547196 6 = 0 |
| data | 0 | 8 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 8 x 0.47712125471966 = 3.8169700377573 | = 0 x 0.4771212547196 6 = 0 |
| way | 0 | 2 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 2 x 0.47712125471966 = 0.95424250943932 | = 0 x 0.4771212547196 6 = 0 |
| amount | 0 | 2 | 0 | 1 | log (3/1) = 0.47712125471966 | = 0 x 0.47712125471966 = 0 | = 2 x 0.47712125471966 = 0.95424250943932 | = 0 x 0.4771212547196 6 = 0 |
| WORD | Q | D1 | D2 | DF | IDF | QDF | QD1 | QD2 |

Description Table TF / IDF:
Password = Term Review of the document
TF = set of documents
Q = Query (Data were tested)
D1 = Data 1
D2 = Data 2
DF = Many of the documents that contain the search terms
IDF = inverse document frequency (log (N / DF))
N = Total documents
DF = Many of the documents containing the search terms.
WDT = Looking weights to calculate the result of multiplying said existing word in the document (TF) to the number of document frequency (IDF).

## H. cosine Similarity

After performing weighting using the term document against tfidf calculation, the final step being made to locate documents relevant to the query is calculating the similarity between two documents by using the formula cosine similarity.

Following the calculation of the cosine similarity:

**Table 6.**
cosine Similarity

| QDFxQD1 | QDFxQD2 | QDF2 | QD12 | QD22 |
|---|---|---|---|---|
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |
| 0.031008131515815 | 0 | 0.031008131515815 | 0.031008131515815 | 0 |
| 0.031008131515815 | 0 | 0.031008131515815 | 0.031008131515815 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0.22764469170526 | 0 | 0 |

| QDFxQD1 | QDFxQD2 | QDF2 | QD12 | QD22 |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0.031008131515815 | 0 | 0.031008131515815 | 0.031008131515815 | 0 |
| 0.031008131515815 | 0 | 0.031008131515815 | 0.031008131515815 | 0 |
| 0 | 0 | 0 | 0.22764469170526 | 0 |
| 0 | 0 | 0 | 0.22764469170526 | 0 |
| 0 | 0 | 0 | 0.22764469170526 | 0 |
| 0 | 0 | 0 | 0.91057876682106 | 0 |
| 0 | 0 | 0 | 0.22764469170526 | 0 |
| 0 | 0 | 0 | 0.22764469170526 | 0 |
| 0 | 0 | 0 | 14.569260269137 | 0 |
| 0 | 0 | 0 | 0.91057876682106 | 0 |
| 0 | 0 | 0 | 0.91057876682106 | 0 |
| 0.12403252606326 | 0 | 1.9451900597054 | 18.56325255419 | 0 |

$$Cos(): \frac{\sum(Qdf\,.Qd1)}{\sqrt{\sum Q_{df}{}^{2}}\,\sqrt{\sum Qd1^{2}}} \text{ ......................................... (2)}$$

Description Formula:
Cos () =
To calculate the similarity vector (document) query Q with each document.
Σ (Qdf. Qd1) =
Calculating the multiplication of Q and other documents. The result is a multiplication of each document with Q summed (corresponding numerator in the formula above).
(Σ Qdf2) 0.5. (Σ Qdf12) 0.5 =
Calculate the length of each document, including Q. How squared weighting each term (word) in each document, add up the value of the square and the last akarkan.

**Table** 7
Status

| Data Answer | weights Similarities | | |
|---|---|---|---|
| | Active | quite Active | Not active |
| 0-5 | <20% | <40% | > 40% |
| 0-10 | <40% | <50% | > 50% |
| 0-15 | <50% | <60% | > 60% |

## I. Result

a. Cos (Q, 1): 0.12403252606326 / (1,94519005910540.5 x 18,563252554190.5) = 2.0640847788532%
b. Cos (Q, 2): 0 / (1,94519005970540.5 x 00.5) = 100%

Based on the results of cosine similarity calculation above, the result of the similarity between the query and d1 is 2.0640847788532% while the results of the query and d2 equation is 100% it can be concluded that the response data is closest to the query response data 2 (d2).
With the results of the data 3 is not active in responding to question on-Moodle LMS.

## 4. Conclusion

Based on the problems that have been pointed out, the conclusions of this study is that the method of term frequency inverse document frequency (tf-idf) and the cosine similarity has been successfully applied in systems with both that the system can deliver the output in the form of an answer that has a value of precise accuracy, with comparison between the answers of each student. By going through the process step of folding case, tokinizing, stopwords and stemming.

So the actor user / admin and faculty can judge by looking at the activity of students in responding to a question on-Moodle LMS.

## 5. Reference

[1] Dwi Smaradahana, Budi Santoso “Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia,” Institut Teknologi Sepuluh Nopember (ITS), Vol. 6, No. 2, 2017.

[2] Jian-Hong Jiang, Rui-Yun MA “*A Code Classification Method Based On TF-IDF*” Guilin University of Electronic Technology, ISBN: 978-1-60595-552-0, 2018.

[3] Muhammad Andi Al-Rizki, Galih Wasis Wicaksono, Yufis Azhar “*The Analysis Of Proximity Between Subjects Based On Primary Contents Using Cosine Similarity On Lective*” Universitas Muhammadiyah Malang, Vol. 2, No. 4, November 2017.

[4] Mali Fauzi, Djoko Cahyo Utomo, Budi Darma Setiawan, Eko Sakti Pramukantoro. “*Automatic Essay Scoring System Using N-Gram And Cosine Similarity For Gamification Based Elearning*” Universitas Brawijaya, Agustus 2017.

[5] Syed Hafeez, Balkrishna Patil “Using explicit semantic similarity for an improved web explorer with ontology and TF-IDF” Everest Education Society’s College of Engineering and technology, Aurangabad, India, Vol 2, issue 7, Jan 2017

[6] Maedeh Afzali, and Suresh Kumar “*Comparative Analysis Of Various Similarity Measure For Finding Similarity Of Two Documents*” Manav Rachna International University, india. Vol.10 No.2 (2017).

[7] Syahroni Wahyu Iriananda, Muhammad Aziz Muslim, Harry Soekotjo Dachlan “Identifikasi Kemiripan Teks Menggunakan Class Indexing Based dan Cosine Similarity Untuk Klasifikasi Dokumen Pengadua” Volume 10, No. 2 (2018).

[8] Rizki Tri Wahyuni, Dhidik Prastiyanto, Eko Suprapto “Penerapan Algoritma *Cosine Similarity* dan Pembobotan TF-IDF pada sistem Klasifikasi Dokumen Skripsi” Universitas Negeri Semarang, Januari-Juni 2017 Vol. 9 No.1.

[9] Rhevitta Widyaning Palupi, Yuita Arum Sari, Putra Pandu Adikara. “Prediksi Rating Novel Baru Berdasarkan Sinopsis Menggunakan Genre Based Collaborative Filtering dan Text Similarity” Universitas Brawijaya, Vol. 3, No. 3, Maret 2019

[10] Rito Putriwana Pratama, Muhammad Faisal, Ajib Hanani, “Deteksi Plagiarisme pada Artikel Jurnal Menggunakan Metode Cosine Similarity” Universitas negeri islam maulana malik ibrahim Malang, Vol.5 No. 1 2019.

[11] Ria Melita, Victor Amrizal, Hendra Bayu Suseno, Taslimun Dirjan. “Penerapan Metode Term Frequency Inverse Document Frequency (tf-idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadist Berbasis Web (Studi Kasus : Syarah Umdatil Ahkam” Universitas Islam Negeri Syarif Hidayatullah Jakarta, VOL 11 NO. 2, Oktober 2018

[12] Ni Luh Ratniasih, Made Sudarma, Nyoman Gunantara. “Penerapan Text Mining dalam Spam Filtering untuk Aplikasi Chat” Teknologi Elektro, Vol. 16, No. 3,September - Desember 2017.

[13] Fauzi Bayu Sejati, Purwono Hendradi, Bambang Pujiarto. “Deteksi Plagiarisme Karya Ilmiah Dengan Pemanfaatan Daftar Pustaka Dalam Pencarian Kemiripan Tema Menggunakan Metode Cosine Similarity” Universitas muhammadiyah magelang, Vol. 2 No. 2 | Januari 2019.

[14] Hartanto “Text mining dan sentimen analisis twitter pada gerakan LGBT”

[15] Ervita Kusuma Putri, Tedy Setiadi “Penerapan text mining pada sistem klasifikasi email spam menggunakan naive bayes”