E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 04, Nomor 02, Desember 2019 : 156 - 165

MANHATTAN DISTANCE AND DICE SIMILARITY EVALUATION ON INDONESIAN ESSAY EXAMINATION SYSTEM

Muhammad Haidar Ali [1], Faisal Rahutomo[2]
[1, 2, 3]Information Technology Department of State Polytechnic of Malang
Jl. Soekarno Hatta No.9, Jatimulyo, Malang
e-mail: haidariyyah96@gmail.com[1], faisal.polinema@gmail.com [2]

ABSTRAK

Setiap proses pembelajaran membutuhkan alat evaluasi untuk mengukur tingkat pemahaman siswa. Jenis evaluasi dapat berupa pertanyaan pilihan ganda, entri pendek dan esai. Beberapa penelitian mengungkapkan esai lebih baik daripada jenis evaluasi lainnya. Penilaian esai secara otomatis diperlukan untuk menghemat waktu guru dalam mengoreksi jawaban. Namun, pengembangan penilaian esai masih berlangsung. Tujuannya adalah untuk mendapatkan nilai akurasi yang lebih baik daripada metode yang digunakan dalam penilaian. Berdasarkan masalah ini, penelitian ini mengusulkan analisis komparatif metode kesamaan untuk penilaian ujian esai online. Metode kesamaan dibandingkan adalah Dice Similarity dan Manhattan Distance. Kedua metode menghasilkan nilai koefisien yang kemudian dibandingkan dengan penilaian sistem dengan skala manual dengan skala yang sama. Data yang digunakan adalah 2162 data. Data ini diperoleh dari 50 siswa yang menjawab 40 pertanyaan (politik, olahraga, gaya hidup, dan teknologi). Data yang diperoleh dalam penelitian ini dapat digunakan untuk mendukung penelitian lain yang dapat diakses di www.indonesian-ir.org. Penelitian ini menunjukkan bahwa skema kemiripan Dice lebih akurat daripada Manhattan Distance.

***Kata Kunci**: Ujian Esai Online, Dice Similarity, Manhattan Distance.*

ABSTRACT

Each learning process requires an evaluation tool to measure the level of understanding of students. The type of evaluation can be multiple choice questions, short entries and essays. Some studies reveal essay exams better than other types of evaluations. An essay assessment is automatically needed to save teacher time in correcting answers. However, the development of essay assessments is still ongoing. The aim is to obtain a better accuracy value than the method used in the assessment. Based on these problems, this study proposes a comparative analysis of similarity methods for online essay exam assessment. The similarity method compared is Similarity Dice and Manhattan Distance. Both methods produce coefficient values which are then compared to the assessment of the system with manual scales with the same scale. The data used were 2162 data. This data was obtained from 50 students who answered 40 questions (politics, sports, lifestyle and technology). The data obtained in this study can be used to support other research that can be accessed at www.indonesian-ir.org. This research shows that the Dice similarity scheme is more accurate than Manhattan Distance.

***Keywords**: Online Essay Assessment, Dice Similarity, Manhattan Distance.*

I. INTRODUCTION

THE development of information technology that is so advanced has helped many people in all fields. One of them is education. This technology overcomes the limitations of time and space in conventional learning. Methods are also widely developed in the learning and supporting technology side. The development of information technology that is increasingly rapid in the era of globalization is now unavoidable influence on the world of education. Global demands require the world of education to always and constantly adjust technological developments towards efforts to improve the quality of education, especially adjusting the use of information and communication technology for the world of education, especially in the learning process.

Every learning process requires an evaluation tool to measure the level of understanding of students. Many types of evaluations, ranging from multiple choice questions, short entries to essays. Some studies reveal that MCQs and short entries are inadequate in the teaching and learning process. Conversely, essay exams can train the delivery of information verbally, this test also requires a better understanding. So that the assessment in essay questions can measure the level of understanding more deeply [1].

Many of the benefits that can be obtained from essay assessments are automatic compared to traditional assessments. In British records, teachers spend 30% of their time correcting student answers and eliminating around 30 billion pounds a year because of this [2]. So that it can be imagined the benefits if an educational institution has a system for automatic assessment especially for essays.

At present, there are many e-learning developments for multiple-choice exam assessments, short entries and essays. However, the development of essay assessments is still ongoing. The aim is to obtain better accuracy in the assessment. This is because the number of methods in stating the suitability of students' answers to the answer keys that have been provided by the teacher [3]. Unfortunately, there has not been an analysis of the comparison of the methods (schemes) that are widely used today.

Some researches on Indonesian essay examinations are Indonesian Language Essay Assessment Using the SVM-LSA Method with Generic Features that produce accuracy of 73%, An Automatic Scoring Tool of Short Text Answer in Indonesian in its application has a standard deviation of 3 -30 of various types of data tested [4].

So that it underlies the absence of a comparison between the vectors of similarity that exist, even though this is very important in determining the method used in making an online essay examination assessment system. In addition, the data used in previous research still uses small size data (classes with fewer respondents and not many types of questions).

Based on this, this study aims to determine the comparison of similarity schemes in Cosine Similarity, Euclidean Distance and Jaccard using 4 problem fields (each question area has 10 questions) with 50 students. Where data obtained from this research will be provided by researchers for other research purposes. On the other hand, the use of stemming in the world of word processing, especially the results of online essay examination, has never been matched, therefore, this study will reveal differences in error values (differences in manual values with system values) using stemming.

II. Literature Review

Literature and learning sources in this research activity use a variety of methods to support the success of system performance, including:

A. Vector Space Model

Vector Space Model is a model used to model a text document as a transformation that can convert digital texts into a more efficient and understandable model so that the analysis process can be carried out. For example, V_d is a vector of documents d. The vector has features in the form of values or weights from the terms in the document [5].

To avoid vectors with large but not important features, the word used as a feature is only when it appears on the training data at least three times, or if the word is not stop-word. The following is the vector space model illustration as shown in Figure 1.

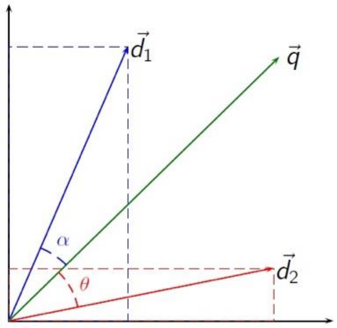


Fig. 1. Vector space model illustration. The vector has features in the form of values or weights from the terms in the document [5].

In VSM, a collection of documents is represented as a term-document matrix (or matrix term frequency). Each cell in the matrix corresponds to the weight given from a term in the specified document. The documents taken are sorted in a sequence that has similarities, the vector model takes into consideration the documents that are relevant to the user's request [6]. Term of documents and its equation can be seen in Figure 2 dan Equation 1.

$$\begin{bmatrix} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{bmatrix}$$

Fig. 2. Term documents example inside Matrix

$$Vd = \{t_1, t_2, .., t_n\} \tag{1}$$

B. Dice Similarity

Where |X| and |Y| are the cardinalities of the two sets (i.e. the number of elements in each set) [7]. The Sørensen index equals twice the n umber of elements common to both sets divided by the sum of the number of elements in each set.

$$DiceSim = \frac{2\,|A\ \cap\ B|}{|A|\ + |B|} \tag{2}$$

C. Manhattan Distance

The next distance metric being used in this research is Manhattan distance. Manhattan distance calculates the distance of two vectors by Equation 4. As same as Euclidean, w_{Ak} is term weighting k in document A and w_{Bk} is term weighting k in document B. The result value is possible to be more than 1 [4].

$$ManhattanDist = \sum_{k=1}^{n} |W_{Ak} - W_{Bk}| \tag{3}$$

D. Nazief Adriani Stemming Algorithm

Nazief andriani algorithm is a special stemming algorithm for Indonesian. This algorithm uses several morphological rules to eliminate affix (prefix, affix, etc.) from a word and then match it in the dictionary of root words (basic words) [8]. So the main basis of this algorithm is a list of basic words. The first step is to collect a list of basic words in Indonesian. The more complete the list, the higher the accuracy of this algorithm.

E. Error Rate

Each result of the implementation of two methods of symmetry metrics will be calculated as percentage error and absolute error. The percentage error value shows how much the difference between the measurement and the fact value. A small error value indicates that the error rate of the system is getting better [2]. Here's the equation of it.

$$Error\ Percentage\ (\%) = \frac{Manual\ Scoring\ \ - System\ Scoring\ (Methods)}{Manual\ Scoring} \tag{4}$$

III. RESEARCH METHOD

The research method is divided into several phases. The phase in detail will be explained in what will be attached below. The figure shows the analysis scheme in the study. The problem faced is that the time is drained for teachers to correct the essay scores of each student. So that a solution is needed to simplify and cut the time of the instructor in providing value.

The first phase is to make a question and answer key, where each question and key has a category. The categories in the exam system are divided into 5, namely lifestyle, sports, politics, economics, and technology. In this phase, the person in charge of the exam question bank is the admin and instructor. Admin and instructor are able to manipulate or make changes to the questions and key answers to the system. The second phase is students start answering exam questions with an online essay scoring system. Each student is able to choose the category of exam questions to be answered. Each category contains 10 (ten) essay questions that students can answer. The more questions answered correctly, the higher the student's score. The third phase is the process of student answer data that will be assessed by the system. System appraisal is done by doing the stages of text-preprocessing and calculation using the Dice similarity scheme and the comparison of the distance of Manhattan [4].

The evaluation process of this system uses a representation of a vector from a student's test answer document and where each component refers to a term. Then the value of each component is the number of occurrences of the term in a document. Once the document is represented as a vector, various vector operations can be performed.

The fourth phase is giving the manual value of students' answers by the system that has been set in the form of absolute numbers and then taken the average value [9]. The fifth phase is calculating the percentage error value between the average manual rating and the rating of the system [10]. Based on this phase, the error value of each method can be obtained. The following is the flow of the online essay exam scoring system as shown in Figure 3.

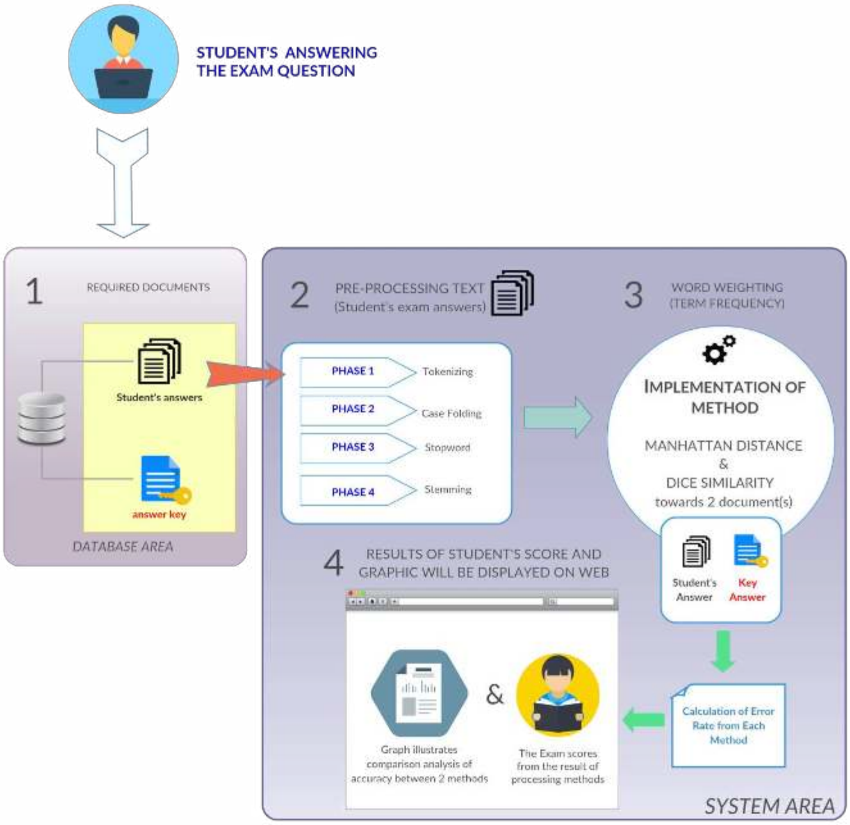

STUDENT'S ANSWERING THE EXAM QUESTION
1
REQUIRED DOCUMENTS
Student's answers
answer key
DATABASE AREA
2
PRE-PROCESSING TEXT (Student's exam answers)
PHASE 1
Tokenizing
PHASE 2
Case Folding
PHASE 3
Stopword
PHASE 4
Stemming
3
WORD WEIGHTING (TERM FREQUENCY)
IMPLEMENTATION OF METHOD
MANHATTAN DISTANCE & DICE SIMILARITY towards 2 document(s)
Student's Answer
Key Answer
Calculation of Error Rate from Each Method
4
RESULTS OF STUDENT'S SCORE AND GRAPHIC WILL BE DISPLAYED ON WEB
Graph illustrates comparison analysis of accuracy between 2 methods
&
The Exam scores from the result of processing methods
SYSTEM AREA

Fig. 3. Research Method

A. Data Source

The data which is needed to build this system is derived from a comprehensive dataset in the form of 40 questions where each 10 questions have different categories (Lifestyle, Politics, Sport, Technology). The answers collected were as many as 2162 (two thousand one hundred and sixty two) texts if they were calculated in their entirety.

Furthermore, the key answers and answers of the students will be processed by the text pre-processing process and the two Manhattan Distance methods and Dice Similarity as a comparison of accuracy.

B. Method Steps

There are stages of pre-processing, which convert text data into numerical data that can be processed. This stage is a very important step before starting the automatic valuation process because this process can affect the accuracy of the assessment [11]. In pre-processing there are several steps that must be done.

In this study the pre-processing stage is using stemming. This is related to the absence of studies showing that stemming use makes judgments more effective [3].

1) Text Pre-Processing

Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and information retrieval (IR) [12]. In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data [13]. The following are the steps that must be carried out on this system.

2) Case Folding

Case folding is a step in text mining to convert uppercase letters to lowercase letters, in the sense that all letters are equaled [14].

TABLE I
AN EXAMPLE OF CASE FOLDING

Before	After
An Ultimate Freedom	an ultimate freedom
Big Body	big body
Nice Jacket	nice jacket

3) Tokenizing

Tokenization is the process of cutting an entire sequence of characters into one word chunk [15].

TABLE II
AN EXAMPLE OF TOKENIZING

I have nothing in this world but you							
⇓ (turn into)							
i	have	nothing	in	this	world	but	you

4) Stopword

The process carried out at this stage is to remove stop-word. Stop-word is a word that is not a unique word in an article or general words that are usually always in an article. Examples of Indonesian words including stop-word are "yang", "dan", "di", "dari", and others [16].

C. Calculation Example

This sub-section describes the calculation example of Dice Similarity and Manhattan Distance. The example used in this study uses the same example as in Ahmad Hafidh Ayatullah's text mining journal [17]. Sentence A is “Lelaki berjenggot itu sedang menggunting kertasnya”. And the sentence B is “pria sedang menggunting kertas”. The processes are described as follows:

1) Case Folding

- Sentence A becomes “lelaki berjenggot itu sedang menggunting kertasnya”.
- Sentence B becomes “pria sedang menggunting kertas”.

2) Tokenizing

- Sentence A becomes an array of words:
{lelaki, berjenggot, itu, sedang, menggunting, kertasnya}.
- Sentence B becomes an array of words:
{pria, sedang, menggunting, kertas}.

3) Stop Words

- Through stop words process, sentence A becomes an array of words:

{lelaki, berjenggot, menggunting, kertasnya}.
- Through stop words process, sentence B becomes an array of words:
{pria, menggunting, kertas}.

4) *Stemming*
- Through stop words process with stemming on it, sentence A becomes an array of words:
{lelaki, jenggot, gunting, kertas}.
- Through stop words process with stemming on it, sentence B becomes an array of words:
{pria, gunting, kertas}.

5) *Term Frequency*
Table 3 describes the results of this step.

TABLE III
PREPROCESSING RESULTS OF SENTENCES A AND B

	lelaki	jenggot	gunting	kertas	pria
A	1	1	1	1	1
B	0	0	1	1	0

6) *Manhattan Distance*
$Manhattan(A,B)$ =|1−0|+|1−0|+|1−1|+|1−1|+|1−0| = 3

7) *Dice Similarity*
A= {lelaki, jenggot, gunting, kertas}
B= {pria, gunting, kertas}
$A \cap B$= {gunting, kertas}
$|A \cap B| = 2$
$|A| = 4$
$|B| = 3$

$$DiceSim = \frac{2\,|A\ \cap\ B|}{|A|\ +|B|} = \frac{2 \times 2}{4+3} = \frac{4}{7} = 0{,}58$$

IV. SOFTWARE IMPLEMENTATION

The application was developed in a web-based environment. MySQL was chosen as the database management system. Then PHP framework Laravel is used to build the application. The students can log in then choose the exam questions to answer based on the available exam questions categories. Each student is required to answer 10 questions for each category they choose. The categories available in this online essay examination system are Life-style, Sports, Politics, and Economics. On the other hand, this system also has role administrators and lecturers page to manage exam questions and answer keys.

A. Database Structure

Figure 4 describes the Physical Data Model (PDM) of "dbessay" database. There are 12 tables in the database with different functions. Each table function is described as follows on Table 4.

As shown in the previous table, 3 main tables for the login function, namely the admins table, student_list, and lecturer. Meanwhile, to save the stages of the text pre-processing process the students' answers using 3 tables, namely answers_of_student, doc_preprocessing_answers_ofstudent, doc_preprocessing_key_answers. The exam_questions and exam_category tables contain question bank data and related question types. The score table is used to store all grades of students who have done the exam questions. Relationships between tables can be seen in Figure 4.

TABLE IV
TABLES NAME WITH THEIR FUNCTIONS

Table(s) name	Function(s)
admins	to store admin users
student_list	to store users list as Student
doc_preprocessing_answers_ofstudent	to save the results of the preprocessing process from student answers
doc_preprocessing_key_answers	to save the results of the preprocessing process from the answer key
answers_of_student	to save all the results of student's exam answers after the process of entering answers.
exam_category	saves the names of exam category
score	to store the Manhattan values
lecturer	to store lecturer users
exam_questions	to save all essay exam questions

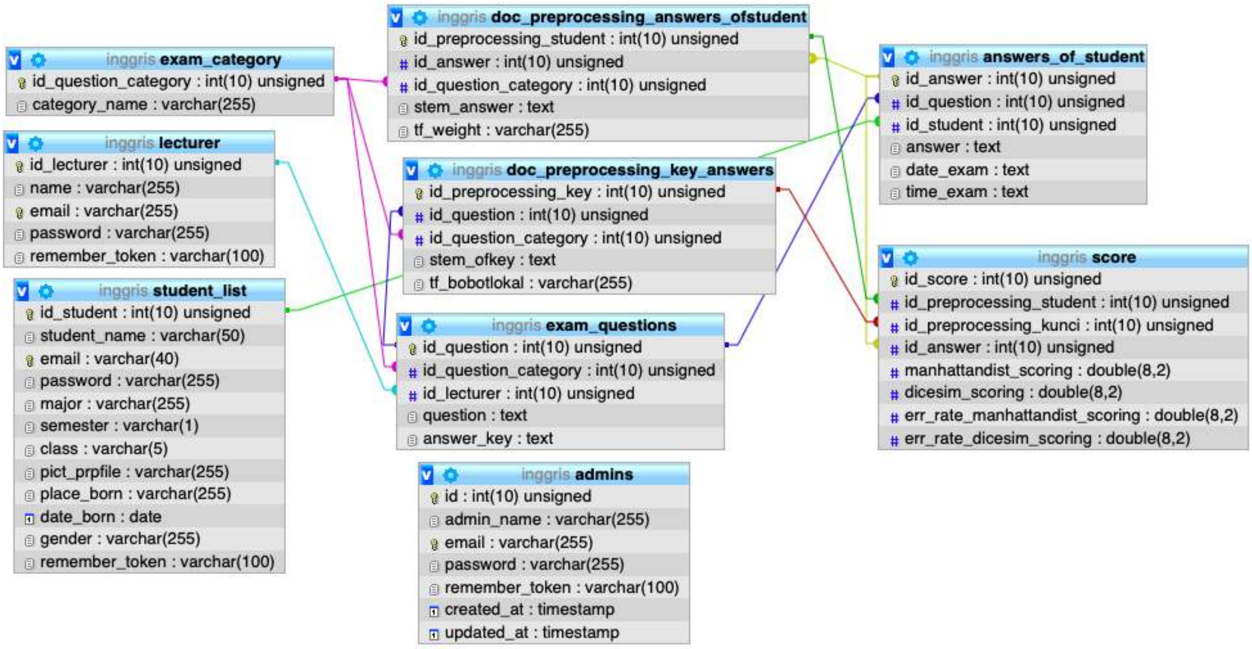

inggris exam_category
id_question_category : int(10) unsigned
category_name : varchar(255)
inggris lecturer
id_lecturer : int(10) unsigned
name : varchar(255)
email : varchar(255)
password : varchar(255)
remember_token : varchar(100)
inggris student_list
id_student : int(10) unsigned
student_name : varchar(50)
email : varchar(40)
password : varchar(255)
major : varchar(255)
semester : varchar(1)
class : varchar(5)
pict_prpfile : varchar(255)
place_born : varchar(255)
date_born : date
gender : varchar(255)
remember_token : varchar(100)
inggris doc_preprocessing_answers_ofstudent
id_preprocessing_student : int(10) unsigned
id_answer : int(10) unsigned
id_question_category : int(10) unsigned
stem_answer : text
tf_weight : varchar(255)
inggris doc_preprocessing_key_answers
id_preprocessing_key : int(10) unsigned
id_question : int(10) unsigned
id_question_category : int(10) unsigned
stem_ofkey : text
tf_bobotlokal : varchar(255)
inggris exam_questions
id_question : int(10) unsigned
id_question_category : int(10) unsigned
id_lecturer : int(10) unsigned
question : text
answer_key : text
inggris admins
id : int(10) unsigned
admin_name : varchar(255)
email : varchar(255)
password : varchar(255)
remember_token : varchar(100)
created_at : timestamp
updated_at : timestamp
inggris answers_of_student
id_answer : int(10) unsigned
id_question : int(10) unsigned
id_student : int(10) unsigned
answer : text
date_exam : text
time_exam : text
inggris score
id_score : int(10) unsigned
id_preprocessing_student : int(10) unsigned
id_preprocessing_kunci : int(10) unsigned
id_answer : int(10) unsigned
manhattandist_scoring : double(8,2)
dicesim_scoring : double(8,2)
err_rate_manhattandist_scoring : double(8,2)
err_rate_dicesim_scoring : double(8,2)

Fig. 3. Physical Data Model

B. *User Interface*

This page is displayed as the student homepage when the application starts before the application enters the test page. The purpose of this page is to introduce users to this application created by the Polinema agency and created with Laravel Framework. This application's home page is shown in Figure 5.

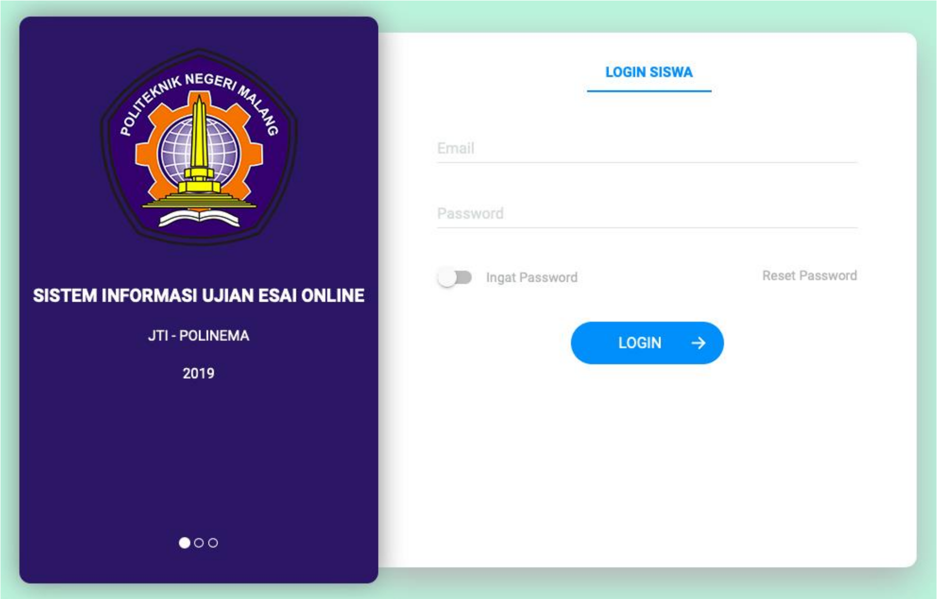
Figure 6 shows an example of answer page for the Lifestyle category exam questions. The number of questions that must be done is 10 items, as stated earlier, each test category has 10 items so that the total items available in this system amount to 40. After the students confirm all of exam answer, the system will automatically correct the students' answers through the stages which are described previously.

V. RESULT AND DISCUSSION

This research is made to improve and develop previous research entitled "Analysis of Aspects of Indonesian Online Language Essay Exams" by Trisna Ari Roshinta [2]. The data used by the two research titles are the same. The data was obtained from 50 students who answered 40 questions (politics, sports, lifestyle and technology).

The data obtained on previous research shows that the percentage error value of Jaccard was 52.31%, Euclidean Distance 332.90%, and Jaccard 59.49%. The following is a comparison of the Error Rate value of the results of the second method of research that has been done (the comparison method is marked by a green column).

After comparing the students' values from the 2 methods of the researchers to the 3 methods of the previous researchers, it can be concluded that the Dice Similarity method has succeeded in obtaining the smallest average error rate, among others, 33.7%. While the highest error rate is generated by the Euclidean method of 333%. Other method error rate results are, Manhattan Distance of 55.4%, cosine of 59.5%, and Jaccard 52.3%. All of the above student grades have surpassed the Stemming process as a refinement of the text pre-processing stage.



SISTEM INFORMASI UJIAN ESAI ONLINE
JTI - POLINEMA
2019

LOGIN SISWA

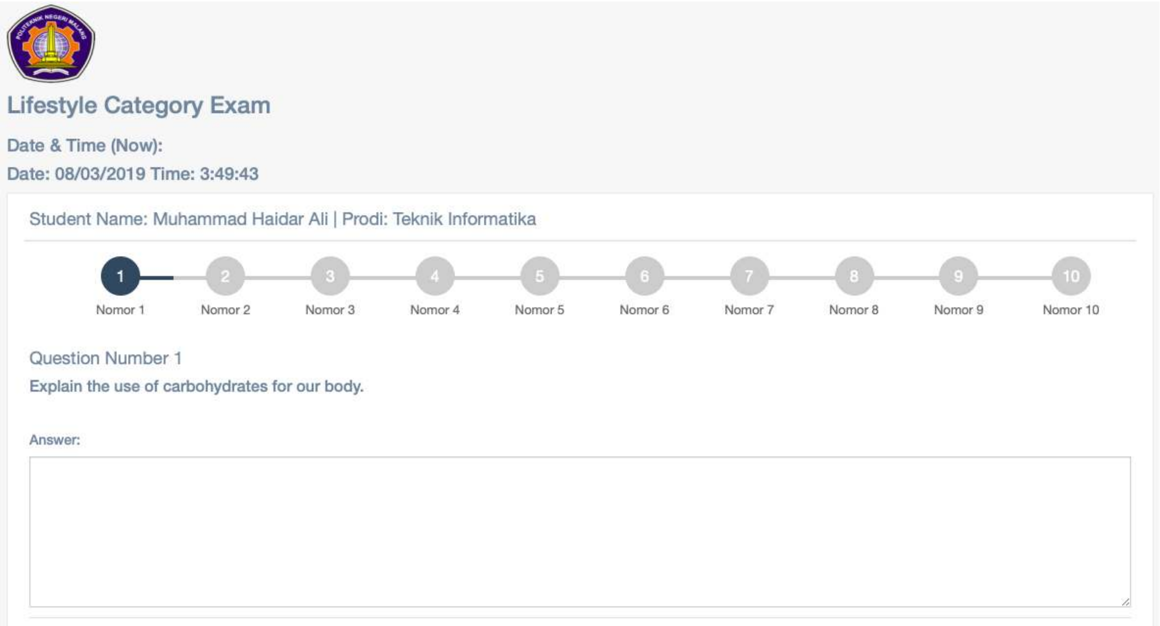
Email

Password

Ingat Password

Reset Password

Fig. 4. Login Screen for Student



Lifestyle Category Exam

Date & Time (Now):
Date: 08/03/2019 Time: 3:49:43

Student Name: Muhammad Haidar Ali | Prodi: Teknik Informatika

Question Number 1
Explain the use of carbohydrates for our body.

Answer:

Fig. 6. Main Page for Essay Exam

TABLE V
TABLE OF COMPARISON AVERAGE OF BOTH RESEARCHES

Category	**Err. Rate Dice (%)**	**Err. Rate Manhattan (%)**	**Error Rate Cosine (%)**	**Error Rate Euclidean (%)**	**Error Rate Jaccard (%)**
Lifestyle	16.08	3.32	54.33131	282.7671	51.31268
	4.9	39.39	52.77894	483.8979	61.19562
	100	4.49	81.22259	414.5803	70.81366
	5.41	15.25	58.97989	415.9395	77.22115
	4.76	4.76	65.44736	220.2017	54.48611
	5.63	10.96	143.9215	629.9405	57.90799
	4.76	4.76	105.53	779.3765	51.85004
	4.76	4.76	76.88124	257.3143	80.24209
	23	0.8	94.79719	255.7896	34.9248
	15	4.59	63.29606	613.8158	72.71217
Politik (Politics)	16.08	3.32	44.93716	202.4219	71.68639
	4.9	39.39	29.18184	73.44606	89.6668
	44	4.49	71.03265	109.6221	17.57667
	5.41	15.25	19.89614	57.1323	35.1428
	4.76	4.76	53.80037	13.8648	60.91255
	5.63	10.96	149.7413	44.22116	48.13708
	4.76	4.76	86.14641	90.6221	21.89834
	4.76	4.76	68.24676	57.13231	60.52151
	23	0.8	71.60619	36.28165	44.93716
	23	4.59	13.8648	89.6668	29.18184
Olahraga (Sport)	10	10	44.22116	68.24676	71.03265
	0.99	11. 31	90.6221	71.60619	19.89614
	0.99	10	57.13231	13.8648	53.80037
	0.99	22	36.28165	44.22116	149.7413
	1.12	10	89.6668	90.6221	86.14641
	0.99	13	25.83338	57.13231	68.24676
	0.99	14	42.19874	36.28165	71.60619
	1.14	10	51.67905	89.6668	13.8648
	0.93	10.38	17.57667	25.83338	44.22116
	0.99	11.31	35.1428	42.19874	90.6221
Teknologi (Technology)	16.08	3.32	60.91255	51.67905	57.13231
	4.9	39.39	48.13708	17.57667	57.13231
	100	4.49	21.89834	48.13708	36.28165
	5.41	15.25	60.52151	21.89834	57.13231
	5	16.08	3.32	60.91255	71.60619
	6.7	4.9	39.39	48.13708	13.8648
	56	100	4.49	21.89834	44.22116
	2.4	5.41	15.25	39.13691	67.66693
	9.3	1.14	10	66.47942	328.2858
	2.5	0.93	11. 31	37.52498	363.9365
Average	**33.65**	**55.40**	**59.48736**	**332.99027**	**52.3078**

VI. CONCLUSION

Based on the results of the analysis, design and implementation carried out, it can be concluded that dice Similarity scheme is known to be more effective than Manhattan Distance. The system method scheme with the smallest percentage error value is the Dice Similarity scheme which is 33.7%. Also, Students who are respondents in this study have varying values that vary, indicating that students work on online essay questions according to their respective abilities. Good student scores on questions with definite types of answers. Because definite answers have a great opportunity to be the same between the key questions and student answers, even without paying attention to the synonyms of the word. From a number of fields taken as test data, it can be seen that the Political problem area is a problem area that has the highest average score compared to the fields of Sports, Technology and Lifestyle.

REFERENCES

[1] N. Suzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "*Automatic Short Answer Grading and Feedback*," pp. 1–20.

[2] T. A. Roshinta and R. Faisal, "*Analisis Aspek-Aspek Ujian Esai Daring Berbahasa Indonesia,*" vol. 01, pp. 1–26, 2016.

[3] T. Dalgleish et al., Text Mining Application and Theory, vol. 136, no. 1. 2007.

[4] T. R. Muzzammil, R. V. H. Ginardi, and D. Purwitasari, "*Modul Klasifikasi Aduan dengan Pendekatan Kemiripan Teks pada Aplikasi Perangkat Bergerak Suara Warga (SURGA) Kota Kediri,*" vol. 5, no. 1, pp. 52–57, 2016.

[5] G. Salton, A. Wong, and C. S. Yang, "*Vector Space Model for Automatic Indexing. Information Retrieval and Language Processing,*" Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.

[6] F. Rahutomo, P. Y. Saputra, C. Febriawan, and P. Putra, "*Implementasi Explicit Semantic Analysis Berbahasa Indonesia Menggunakan Corpus Wikipedia Indonesia,*" J. Inform. Polinema, vol. 4, no. 4, pp. 252–257, 2018.

[7] F. Amin and E. Winarno, "*Rancang Bangun Sistem Temu Kembali Informasi (Information Retrieval System) Dokumen Berbahasa Jawa menggunakan Metode DICE Similarity,*" vol. 21, no. 2, pp. 99–106, 2016.

[8] F. Rahutomo, Z. Hanif, R. Adi, and I. F. Rozi, "*Implementasi Text Mining Pada Laman Blog di*," pp. 101–109, 2018.

[9] V. M.K and K. K, "A Survey on Similarity Measures in Text Mining," Mach. Learn. Appl. An Int. J., vol. 3, no. 1, pp. 19–28, 2016.

[10] R. Feldman and J. Sanger, The Text Mining handbook. .

[11] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "*Automatic Text Scoring Using Neural Networks*," pp. 715–725, 2016.

[12] M. Shoaib, A. Daud, M. Sikandar, and H. Khiyal, "*An Improved Similarity Measure for Text Documents*," J. Basic. Appl. Sci. Res, vol. 4, no. 6, pp. 215–223, 2014.

[13] C. C. Aggrawal and C. Zai, Mining Text Data. .

[14] W. H.Gomaa and A. A. Fahmy, "*A Survey of Text Similarity Approaches*," Int. J. Comput. Appl., vol. 68, no. 13, pp. 13–18, 2013.

[15] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, "*Data mining: A knowledge discovery approach*," Data Min. A Knowl. Discov. Approach, no. September 2017, pp. 1–606, 2007.

[16] M. Astiningrum et al., "*Implementasi nlp dengan konversi kata pada sistem chatbot konsultasi laktasi*," vol. 5, no. November, pp. 46–52, 2018.

[17] F. Rahutomo and A. Hafidh Ayatullah, "*Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis,"* Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 3, no. 4, p. 319, 2018.