ISSN 2085-4552

ULTIMATICS

Jurnal Teknik Informatika

Engineering | Vol. 13 | No. 2 | Page. 72 - 141 | December 2021 | ISSN 2085-4552

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

ISSN 2085-4552

EDITORIAL ADDRESS

Universitas Multimedia Nusantara (UMN)
Jl. Scientia Boulevard
Gading Serpong
Tangerang, Banten - 15811
Indonesia
Phone. (021) 5422 0808
Fax. (021) 5422 0800
Email : ultimatics@umn.ac.id

Ultimatics : Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Computer Science and Informatics, as well as the latest theoretical and practical issues, including Analysis and Design of Algorithm, Software Engineering, System and Network Security, Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics. Ultimatics : Jurnal Teknik Informatika is published regularly twice a year (June and December) and is published by the Faculty of Engineering and Informatics at Universitas Multimedia Nusantara.

Call for Papers

International Journal of New Media Technology (IJNMT) is a scholarly open access, peer-reviewed, and interdisciplinary journal focusing on theories, methods and implementations of new media technology. Topics include, but not limited to digital technology for creative industry, infrastructure technology, computing communication and networking, signal and image processing, intelligent system, control and embedded system, mobile and web based system, and robotics. IJNMT is published twice a year by Faculty of Engineering and Informatics of Universitas Multimedia Nusantara in cooperation with UMN Press.

Ultimatics : Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Analysis and Design of Algorithm, Software Engineering, System and Network security, as well as the latest theoretical and practical issues, including Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics.

Ultima Computing : Jurnal Sistem Komputer is a Journal of Computer Engineering Study Program, Universitas Multimedia Nusantara which presents scientific research articles in the field of Computer Engineering and Electrical Engineering as well as current theoretical and practical issues, including Edge Computing, Internet-of-Things, Embedded Systems, Robotics, Control System, Network and Communication, System Integration, as well as other topics in the field of Computer Engineering and Electrical Engineering.

Ultima InfoSys : Jurnal Ilmu Sistem Informasi is a Journal of Information Systems Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the field of Information Systems, as well as the latest theoretical and practical issues, including database systems, management information systems, system analysis and development, system project management information, programming, mobile information system, and other topics related to Information Systems.

ISSN 2085-4552

FOREWORD

ULTIMA Greetings!

Ultimatics : Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Computer Science and Informatics, as well as the latest theoretical and practical issues, including Analysis and Design of Algorithm, Software Engineering, System and Network Security, Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics. Ultimatics : Jurnal Teknik Informatika is published regularly twice a year (June and December) and is published by the Faculty of Engineering and Informatics at Universitas Multimedia Nusantara.

In this December 2021 edition, Ultimatics enters the 2nd Edition of Volume 13. In this edition there are nine scientific papers from researchers, academics and practitioners in the fields of Computer Science and Informatics. Some of the topics raised in this journal are: Recommendation for Classification of News Categories Using Support Vector Machine Algorithm with SVD, Digital Image Processing using Texture Features Extraction of Local Seeds in Nekbaun Village with Color Moment, Gray Level Co Occurance Matrix, and k-Nearest Neighbor, The Implementation of the Weight Product (WP) Method on the Best Employee Selection, Fasttext Word Embedding and Random Forest Classifier for User Feedback Sentiment Classification in Bahasa Indonesia, Spam Filtering on User Feedback Via Text Classification Using Multinomial Naïve Bayes and TF-IDF, Classification of Metagenome Fragments with Agglomerative Hierarchical Clustering, Elastic Stack Ability Test Monitoring Slowloris Attack on Digital Ocean Server, Information Systems and Recommendations using AHP at SMA Islamic Center Tangerang and Test Case Analysis with Keyword-Driven Testing Approach Using Katalon Studio Tools.

On this occasion we would also like to invite the participation of our dear readers, researchers, academics, and practitioners, in the field of Engineering and Informatics, to submit quality scientific papers to: International Journal of New Media Technology (IJNMT), Ultimatics : Jurnal Teknik Informatics, Ultima Infosys: Journal of Information Systems and Ultima Computing: Journal of Computer Systems. Information regarding writing guidelines and templates, as well as other related information can be obtained through the email address ultimatics@umn.ac.id and the web page of our Journal here.

Finally, we would like to thank all contributors to this December 2021 Edition of Ultimatics. We hope that scientific articles from research in this journal can be useful and contribute to the development of research and science in Indonesia.

December 2021,

M.B.Nugraha, S.T., M.T.
Editor-in-Chief

TABLE OF CONTENT

ISSN 2085-4552

Recommendation for Classification of News Categories Using Support Vector Machine Algorithm with SVD

Nofenky[1], Dionisia Bhisetya Rarasati[2]
Fakultas Teknologi dan Design, Universitas Bunda Mulia, Jakarta, Indonesia
[1]nofenky@gmail.com, [2]rarasati.ubm@gmail.com

Accepted 30 November 2020
Approved 10 August 2021

***Abstract*— online news is a digital information media currently has a very easy and flexible updating process. The News Document grouping process is implemented in several stages, including Text Mining which includes Text Pre-processing which includes Tokenizing, Stopword removal, Stemming, Word Merging, TF-IDF and Confusion Matrix. Of the several techniques in Text Mining, the most frequently used for News Document classification is Support Vector Machine (SVM) Algorithm. SVM has many advantages of being able to identify separate hyperplane that maximizes the margin between two or more different classes. The selection of the features in SVM Algorithm significantly affects the classification accuracy results. Therefore, in this study a combination of the feature selection methods is used, namely Singular Value Decomposition in order to increase accuracy and reduce the Classifier Time Support Vector Machine. This research resulted in text classification in the form of categories Entertainment, Health, Politics and Technology. Based on the Support Vector Machines Algorithm, an accuracy rate of 81% was obtained with 360 Data Training and 120 Data Testing, after adding the Singular Value Decomposition feature with a K-Rank value of 50%, a significant increase in accuracy was obtained with an accuracy value of 94% and The time of Algorithm process is faster.**

***Index Terms*— News Classification, Support Vector Machine, Singular Value Decomposition, Text Mining.**

I. Introduction

Information has become one of the necessities in everyday human life. Information can be interpreted as knowledge from a learning, experience or instruction. In some cases knowledge about events or situations that have been collected or received through the communication process, gathering news from certain events.[1] In the modern era and the sophistication of technology, speed and accuracy in obtaining news or information are needed by the community. Newspapers are the people's choice in obtaining fast and accurate news and information.

In general, the news that is conveyed in news portals consists of various categories such as Political News, Technology News, Entertainment News, Health News and others. For example on the Indonesian news website www.kompas.com, www.waspada.com, www.vivanews.com, etc. [2]

However, in distributing news into certain categories, for now it is still done manually, where the news is separated by the author, meaning that in uploading news the writers must first know what the content of the news they will upload will be and then it will be included in the right and correct category. Therefore, it is necessary to have a system that can classify news categories automatically according to the existing news categories based on the title and content of the news so that this system can assist news uploaders in uploading news.

The SVM algorithm can be explained simply as an attempt to find the best hyperplane that can function as a separator between 2 classes in the input space. The SVM method is rooted in statistical learning theories where this method is very promising in providing better results than some other similar methods. [3]

To be able to produce higher accuracy than a Support Vector Machine Algorithm, an experiment will be carried out by adding the Singular Value Decomposition (SVD) Method, SVD is a matrix decomposition technique to facilitate data processing, because the Singular Value Decomposition (SVD) Algorithm has advantages over process time efficiency Algorithm on a large-scale Dataset and Singular Value Decomposition (SVD) Algorithm were also chosen because they have the ability to perform the decomposition process on a term-document matrix, so that a matrix that still stores important information with far-reaching smaller dimensions can be obtained.

The purpose of this research is to build a system that can automatically classify news categories into a news category into an actual category, to find out the difference in the final results of News Classification with the Support Vector Machine Algorithm only and with the addition of the Singular Value Decomposition Method, and also to make it easier for news uploaders to upload news with an automatic news category categorization system.

II. LITERATURE REVIEW

A. *Text Mining*

Text mining is a technique in computer science that can be used to solve a large number of information problems by combining techniques from Data Mining, Machine Learning, Natural Language Processing, Information Retrieval and Knowledge Management. [4]. Text Mining seeks to extract useful information from data sources through an identification and exploration of interesting patterns. In Text Mining the data source can be obtained from a collection of documents, this means that the data can be in the form of newspapers, magazines, articles, letters, or research reports such as journals, or a thesis.

B. *Stages of Text Mining*

An explanation of each of the Text Preprocessing processes is as follows:[5]

1). Tokenizing.

Tokenizing Text is unstructured data that must be changed first to make it structured before further analysis. The text in the email entered into the application is stored in a 1-dimensional array. The words in a sentence are divided based on the sentence and then the words will be divided again based on spaces.

2). Stopword removal.

After doing the Tokenizing Text process that the word is not tied to other words. As a result of this separation, there are some words that have no relevant meaning at all in determining the characteristics of a tokenized document, such as the words "this, that, and, or" and many more words. kind. Words that have no relevant meaning are called stopwords.

3). Stemming.

Stemming is a process of mapping and parsing in the Variant form of a word into its basic word form. In Indonesian documents the Stemming process is very necessary before entering the Text Mining process because Indonesian has Prefixes, Suffixes, Infixes and Confixes which make a basic word can be changed into many forms and as a result of making word searches difficult. The following are examples and meanings of affixes in Indonesian.

a. *Suffixes* (Akhiran) is an affix that is usually added to the end of a word, such as “-an, -kan, and ‘-i”.

b. Prefixes (Awalan) is an affix that is usually added to the beginning of a root word or the basic form of a word, such as “-per, -mem”

c. *Confixes* (Sifiks and Prefiks) a single affix occurs from two separate elements, such as “ke-....-an”

4). Word weighting.

The TF-IDF method is a method for calculating the weight of each word that is most commonly used in Information Retrieval. This method is also known to be efficient, easy to use and has high accuracy results.[6]

The data that has gone through the preprocessing stage is in the form of a numeric using this method TF-IDF. The Term Frequency Inverse Document Frequency (TF-IDF) method is a method commonly used to determine how far connected a word (term) is to a document by giving weight to each word. The TF-IDF method itself combines two concepts, namely the frequency of occurrence of a word in a document and the inverse frequency of a document containing these words.[7]

In calculating the weight with TF-IDF, what is first calculated is the TF value of a word with the weight of each word being 1. TF (Term Frequency) which states the number of words that appear in a document. DF (Document Frequency) states how many documents for TF calculation using the following formula: [8]

$$D\,F = \log\left(\frac{N}{DF(W)}\right) \quad (1)$$

$$TF - IDF(w, d) = TF(w, d) \times IDF(w) \quad (2)$$

$$IDF(Word) = \log \frac{td}{df} \quad (3)$$

Explanation:

- *TF-IDF*(w,d): the weight of one word in the entire document.
- w: word.
- d: document.
- *TF*(w,d): frequency of occurrence of a word (w) in a document (d).
- *IDF*(w): inverse *DF* from word (w).
- N: total number of documents.
- *DF*(w): the number of documents containing a word (w).

IDF(word) is an IDF value of each word to be searched, while TD is the total number of existing documents and DF is the number of occurrences of words in a document.

C. *Confusion Matrix*

Confusion Matrix is the method used in the calculation of accuracy. In testing the accuracy and search results will be evaluated into the value of Recall, Precision and Accuracy. Where Precision is an evaluation of the ability of the system to find the most relevant ranking and can be defined as a percentage of documents that are retrieved and are truly relevant to the Query. Recall is an evaluation of the system's ability to find all relevant items from collections and can be defined as a percentage of documents relevant to the Query. Meanwhile, Accuracy is a comparison between cases that will be correctly identified and the total number of existing cases.[6]

Table 1. Confusion Matrix.

Document	True Value	
	Relevant	Non Relevant
Retrieve d	True Positive (tp) Correct result	False Positive (fp) Unexpected result
Not Retrieve d	False Negative (fn) Missing Renault	True Negative (tn) Corect absence of Renault

Explanation:

- *TP (True Positive)* = Number of correctly labelled positive samples.
- *FP (False Positive)* = Number of negative samples incorrectly labelled as positive.
- *FN (False Negative)* = Number of positive samples incorrectly labelled as negative
- *TN (True Negative)* = Number of correctly labelled negative samples

formula as follows:

$$precision = \frac{tp}{(tp+fp)} \quad (4)$$

$$Recall = \frac{tp}{(tp+fn)} \quad (5)$$

$$Accuracy = \frac{(tp+tn)}{(tp+fp+tn+fn)} \quad (6)$$

D. *Support Vector Machine*

Support Vector Machine (SVM) is a technique that is relatively new when compared to other existing techniques, but SVM has a much better performance in various application fields such as Bioformatics, Handwriting Recognition, Text Classification and so on. [3]

SVM is a technique for making predictions, both in the case of classification and regression. SVM has the basic principle of linear classifier which means that linear classification cases can be separated, but SVM has been developed so that it can work on non-linear problems by including kernel concepts in a high-dimensional workspace. The kernel function is commonly used to map a lower initial dimension to a relatively higher dimension. Kinds of kernel functions include:[10]

Formula *Support Vector Machine*

$$f(xd) = \sum aiyiK\,(xi\,,\,xd) + \mathrm{b}\,(xisv)\,ns\;i{=}1 (7)$$

Kernel Function

- Kernel linier
  $K(xi, x) = xiTx$ (1)
- *Polynomial*
  $K(xi, x) = (\gamma.\, xiTx + r)p,\ \gamma > 0$ (9)
- Radial basis function
  $K(xi, x) = exp(-\gamma|xi - x|2),\ \gamma > 0$ (10)
- Sigmoid kernel
  $K(xi, x) = tanh(\gamma.\, xiT + r)$ (11)

E. *Singular Value Decomposition (SVD)*

Singular Value Decomposition is one of the many techniques in processing a matrix derived from the science of linear algebra which was introduced by Beltrami in 1873. SVD is one of the stages in the process contained in the Latent Semantic Analysis (LSA) method. used to process a matrix in linear algebra which is used as a tool in mathematics and is commonly used to represent a matrix and is capable of performing various analyzes and matrix computations.[11] SVD is very useful in decomposing a matrix divided into 3 new matrices, including the U orthogonal matrix, the S diagonal matrix and the last one is the transpose matrix of the D orthogonal matrix or it can also be formulated as follows:

Formula *Singular Value Decomposition.*

$$A_{m\,x\,n} = U_{m\,x\,n}\,S_{m\,x\,n}\,{V^t}_{n\,x\,n} \quad (12)$$

Explanation:

- mxn is a matriks
- A Matriks size mxn
- U The singular vector of the matrix A and this vector is orthonormal
- S The diagonal of the vector that composes the singular value of the corresponding singular vector.
- VT The singular vector of matrix A is also orthonormal.

F. *F1-Score*

F1 or F-Measure is a harmonic mean of precision and recall or it can be abbreviated as f1-score which is a comparison of the mean or average of precision and recall that is weighted. The range of an f1 value is 0 to 1. Here's the equation:[11]

$$f1 = 2 \times \frac{P \times R}{P+R} \quad (13)$$

Explanation:

- F1 : F-Measure or F-Score
- P : Precision
- R : Recall

III. RESEARCH METODOLOGY

In this chapter, it is explained about the stages in system design that will be made by the author, which includes Data, System Description, Analysis Model and Interface Design. Here's the explanation

A. Development Method

The method of developing a system used in this research is the Waterfall Method or Structured Approach. In general, the Waterfall method is a method that is often used to analyze systems. The essence of this method is the stages in working on a system that are carried out sequentially, the Waterfall method consists of several stages of activities, including:

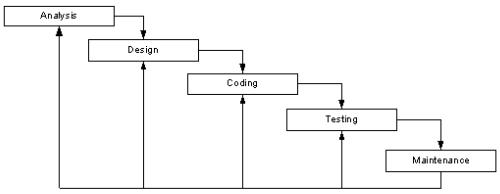


Figure 1. *Development Method waterfall.*

B. Diagram Block

Below is a system process using Block Diagrams:

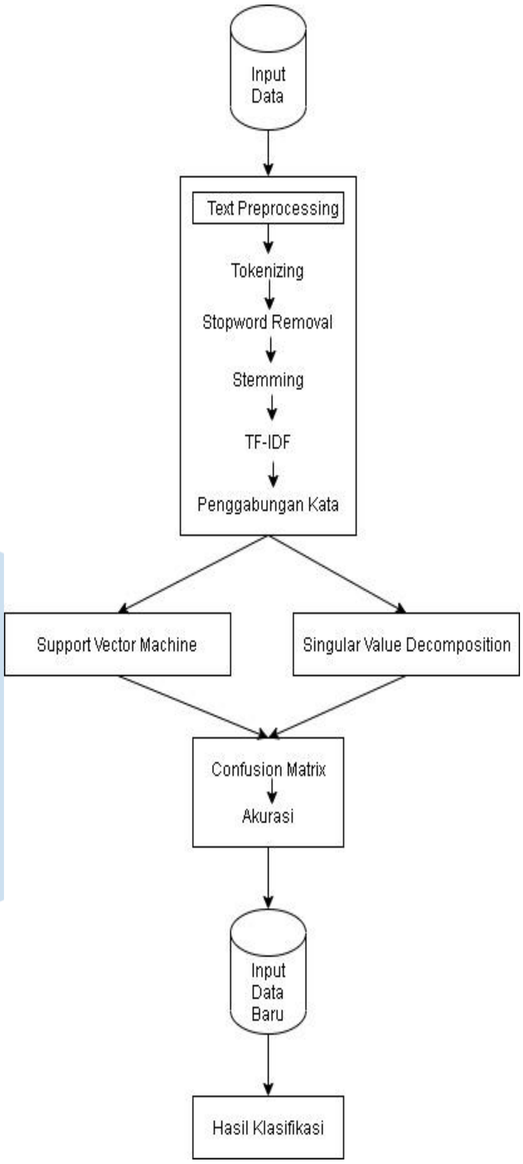


Figure 2. *Diagram Block.*

The classification process starts from input data in the form of news titles, then continues with text operations, in this process there are several stages, namely the tokenizing stage to separate words and convert them into spaces, stopword stages to delete words that do not contain meaning, stemming stages to remove affixed words and weighting or TF-IDF for the process of giving the index or frequency contained in the final word of the stemming process, then it will enter the word merging process (synonym), if there are different words but have the same meaning, then the system can combine them together with the frequency, Stages The next step is Support Vector Machine and Singular Value Decomposition. Next is the testing phase using the Confusion Matrix, the number of correct predictions is divided by the total of all data. And the last one is Classification with new text data in the form of title and news content to determine the category of the news.

C. *Tokenizing*

In the tokenizing the process that occurs is the process of breaking sentences into word for word, these words are also changed from uppercase to lowercase and eliminate unique characters that are not included in the word.

D. *Stopword Removal*

The next process is Stopword, which is the process of filtered words where unimportant words from the text will be discarded. The system will check the Stopword list dictionary, if the word exists then the word will be deleted.

E. *Stemming*

After the Stopword stage, the next process is the Stemming process. Where the system will search for words from the existing news text and convert them into basic words.

IV. RESULT AND DISCUSSION

In this chapter, are the stages of implementation and testing of the system that has been built.

A. *User Interface*

The image below is an image of the Process Data, display from the program that was created.

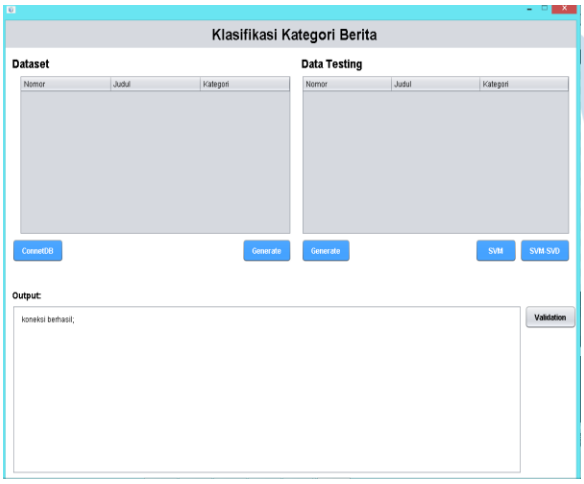


Figure 3. *Process Data.*

News Classification

The image below is a display of the new news classification

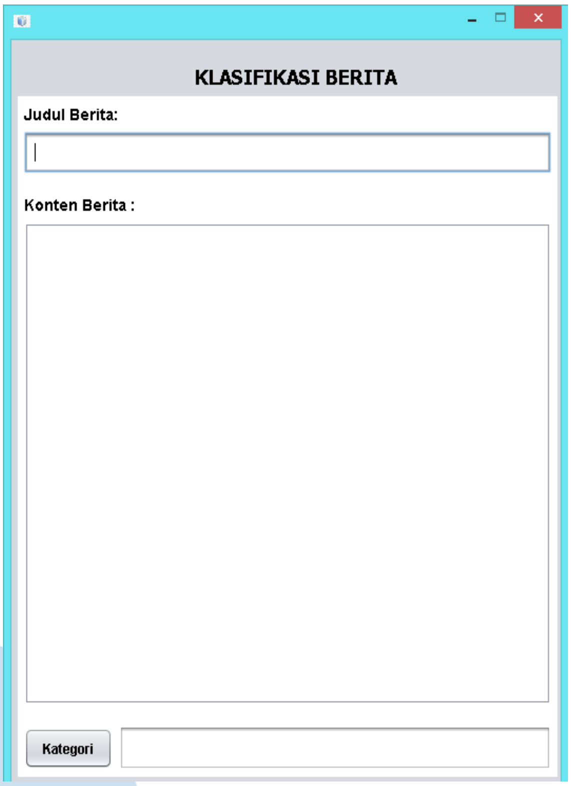


Figure 4. *News Classification*

B. Support Vector Machine Algorithm Classification

In this stage, a news category classification process will be carried out based on the title and content of the news, there are 480 data on the title and content of the news with the division of 360 Training Data and 120 Testing Data, the following results are obtained:

Total Term TF-IDF 4092 from 480 data documents News In this process the system will carry out a machine learning and training process which only uses 1 algorithm, namely Support Vector Machine and accuracy calculations on the Confusion Matrix.

Train Data

```
dataBerita.addAll(train1);
dataBerita.addAll(train2);
kelas[0]=train1.get(0).getKategori();
kelas[1]=train2.get(0).getKategori();
data=new
double[dataBerita.size()][dataBerita.get(0).getSvd(
).length];
y=new int[dataBerita.size()];
alpha=new double[dataBerita.size()];
Arrays.fill(alpha, 0);
deltaAlpha=new double[dataBerita.size()];
Arrays.fill(deltaAlpha, 0);
error=new double[dataBerita.size()];
w=new double[dataBerita.size()];
int i;
for(i=0; i<dataBerita.size(); i++){
data[i]=dataBerita.get(i).getSvd();
if(dataBerita.get(i).getKategori().equals(kelas[0]))
```

```
y[i]=-1;
else y[i]=1; }
setKernel();

Set Kernel
kernel=new double[data.length][data.length];
int i,j;
for(i=0; i<data.length; i++){
for(j=0; j<data.length; j++){
kernel[i][j]=hitungKernel(data[i], data[j]); }}

Hitung Kernel
int row=x1.length;
int a=0;
double hasil=0;
while(a<row){
hasil+=x1[a]*x2[a];
++;}
hasil+=c;
hasil=Math.pow(hasil, d);
return hasil;
```

Figure 5. *Source Code SVM*

```
Output:

Confusion Matrix
Kategori      Entertaiment  Kesehatan  Politik  Teknologi
Entertaiment  14            0          11       2
Kesehatan     0             6          23       2
Politik       0             0          31       0
Teknologi     0             0          8        23

TP:           14            6          31       23
FP:           0             0          42       4
FN:           13            25         0        8
TN:           93            89         47       85

Akurasi:      0,89          0,79       0,65     0,90
Presisi:      1,00          1,00       0,42     0,85
Recall:       0,52          0,19       1,00     0,74

Rata-rata
Akurasi:      0,81
Presisi:      0,82
Recall:       0,61

Waktu Proses:  35,81 detik
---------------------------------------------
```

Figure 6. *Confusion Matrix SVM Algorithm.*

From Figure 5 we get:

- The Accuracy Value for the Entertainment Category is 89%, the Health Category is 79%, the Political Category is 65% and the Technology Category is 90%.
- The Precision Value for the Entertainment Category is 100%, the Health Category is 100%, the Political Category is 42% and the Technology Category is 85%.
- Recall value for Entertainment Category is 52%, Health Category is 19%, Political Category is 100% and Technology Category is 74%.

And Average Score:

- The average value of the Accuracy Value is 81%
- The Average Value of Precision is 82%
- The average value of the recall value is 61%

Calculation f1-*score*

$$f1 = 2 \times \frac{82 \times 61}{82 + 61} = 69{,}96$$

From the calculation of the f1-score against the Support Vector Machine algorithm, a value of 69.96% is obtained, and with the Algorithm processing time for 35.81 seconds.

C. Classification of Support Vector Machine Algorithm with SingularValue Decomposition

Singular Value Decomposition is a technique used to decompose matrices of any size. to facilitate data processing. Here it is used with a K-Rank value of 50% then the results are as follows:

```
setRank
System.out.println(k);
if(k<1 || k>=Math.min(m, n)) return;
int i, j;
double[][] hasil;
hasil=new double[m][k];
for(i=0; i<m; i++){
for(j=0; j<k; j++){
hasil[i][j]=u[i][j];}
u=hasil;

GetVector
double[] hasil=new double[s.length];
double[] temp=new double[s.length];
int i, j;
for(i=0; i<u[0].length; i++){
for(j=0; j<tfidf.length; j++){
temp[i]+=tfidf[j]*u[j][i]; } }
for(i=0; i<si.length; i++){
for(j=0; j<temp.length; j++){
hasil[i]+=temp[j]*si[j][i]; } }
return hasil;
u=hasil;
hasil=new double[k][k];
for(i=0; i<k; i++){
for(j=0; j<k; j++){
hasil[i][j]=s[i][j]; }}
s=hasil;
```

```
Matrix msi=new Matrix(s);
si=msi.inverse().getArrayCopy();
hasil=new double[n][k];
for(i=0; i<n; i++){
for(j=0; j<k; j++){
hasil[i][j]=v[i][j]; }}

v=hasil;
```

Figure 7. *Source Code SVD.*

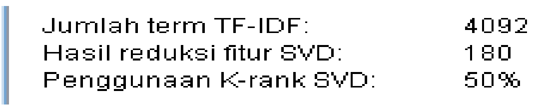
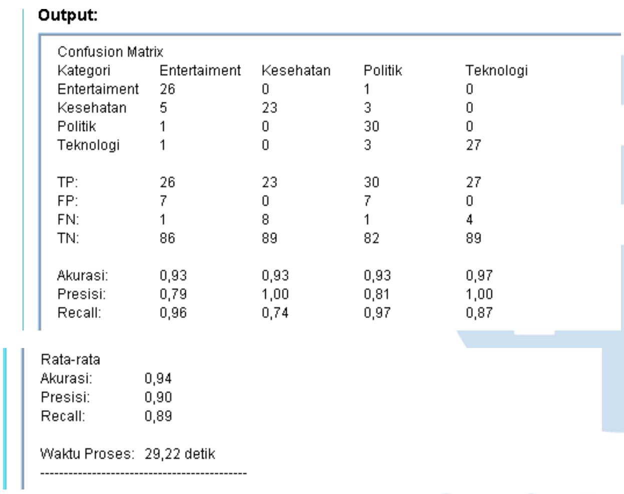

Jumlah term TF-IDF: 4092
Hasil reduksi fitur SVD: 180
Penggunaan K-rank SVD: 50%

Figure 8. Results of Using K-Rank SVD.


Output:

Confusion Matrix

Kategori	Entertaiment	Kesehatan	Politik	Teknologi
Entertaiment	26	0	1	0
Kesehatan	5	23	3	0
Politik	1	0	30	0
Teknologi	1	0	3	27
TP:	26	23	30	27
FP:	7	0	7	0
FN:	1	8	1	4
TN:	86	89	82	89
Akurasi:	0,93	0,93	0,93	0,97
Presisi:	0,79	1,00	0,81	1,00
Recall:	0,96	0,74	0,97	0,87

Rata-rata
Akurasi: 0,94
Presisi: 0,90
Recall: 0,89

Waktu Proses: 29,22 detik

Figure 9. Confusion Matrix Results of SVM and SVD Algorithms.

From Figure 7 we get:

- Accuracy Value for Entertainment Category is 93%, Health Category is 97%, Political Category is 93% and Technology Category is 97%.
- The Precision Value for the Entertainment Category is 79%, the Health Category is 100%, the Political Category is 81% and the Technology Category is 100%.
- Recall value for Entertainment Category is 96%, Health Category is 74%, Political Category is 97% and Technology Category is 87%.

And Average Score:

- The average value of the Accuracy Value is 94%
- Average Precision Value of 90%
- The average value of the recall value is 89%

Calculation f1-*score*

$$f1 = 2 \times \frac{90 \times 89}{90 + 89} = 89{,}50$$

From the calculation of f1-score against the Support Vector Machine algorithm with Singular Value Decomposition, the value is 89.50%, and with the Algorithm processing time for 29.22 seconds.

D. Comparison of SVM and SVM-SVD Hasil Results

Table 2. *Comparison of SVM and SVM-SVD Results*

	Support Vector Machine				
	E	K	P	T	**Average**
Accuracy	89	79	65	90	**81**
Precision	100	100	42	85	**82**
Recall	52	18	100	74	**61**
	Support Vector Machine dan Singular Value Decomposition				
	E	K	P	T	**Average**
Accuracy	93	93	92	97	**94**
Precision	79	100	81	100	**90**
Recall	96	74	97	87	**89**

From table 2, the comparison between the Support Vector Machine Algorithm and the Singular Value Decomposition results in a higher level of Precision, Recall and Accuracy using the Support Vector Machine Algorithm with the addition of the Singular Value Decomposition Method rather than using only one Support Vector Machine Algorithm. And also for the time required for the classification process on the Support Vector Machine Algorithm with the Singular Value decomposition method, the classification duration is faster with a time of 29.22 seconds compared to only using the Support Vector Machine Algorithm with a time of 35.81 seconds.

E. New Data Classification

At this stage, data validation will be carried out by entering New Data in the form of News Title and Contents then the system will automatically detect the News Category

Experiment with new data.

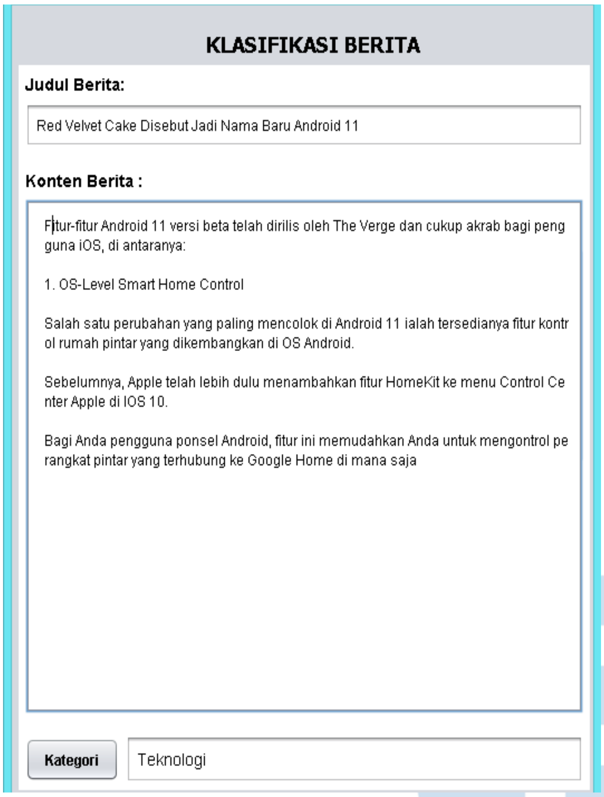


Figure 10. *New Data Classification.*

The results of experiments with News Title and Content are classified as Technology category.

From the results of the classification in Figure 8, it is known the accuracy of each category. The accuracy of each category is obtained from the number of words/terms after passing through the Text Preprocessing stages. The existing words/terms will be checked into the Training Data and Testing Data which will then be calculated for the total Words/Terms contained in each news category in the existing Training Data and Testing Data.

```
Tingkat Persentasi Masing-Masing Kategori

Entertaiment      6,90 %     2 Kata
Kesehatan         6,90 %     2 Kata
Teknologi        75,86 %    22 Kata
Politik          10,34 %     3 Kata
Total            100 %      29 Kata

Berita Tersebut Masuk Kategori Teknologi dengan Akurasi 75,86 %
```

Figure 11. Accuracy of each Category.

V. CONCLUSION

The conclusions that can be drawn from the development of the News classification recommendation system with the Support Vector Machine Algorithm and Singular Value Decomposition are as follows:

1) The accuracy level of the Support Vector Machine Algorithm from Experiments with a total of 360 Training Data and 120 Testing Data obtained an Accuracy Value of 81%
2) The Accuracy Level of the Support Vector Machine Algorithm and Singular Value Decomposition from Experiments with a total of 360 Training Data and 120 Testing Data obtained an Accuracy Value of 94%
3) From the results of experiments conducted Singular Value Decomposition can be applied with a Support Vector Machine and there is an increase in accuracy results and faster processing time.
4) From the results of the f1-score calculation, the Support Vector Machine Algorithm and Singular Value Decomposition values are higher by 89.50% compared to using the Support Vector Machine Algorithm only with a value of 69.96%.

The Support Vector Machine and Singular Value Decomposition Algorithms in this study were only tested with a little data, it would be better if there were more data.

REFERENCES

[1] I. S. Widodo, S. Pd, M. Si, P. Kasih, and M. Kom, “Klasifikasi Konten berita Surat Kabar Berdasarkan Judul Dengan Text Mining Menggunakan Metode Naive Bayes(STUDI KASUS : RADAR KEDIRI) Oleh : Enggal Suci Febriani Dibimbing oleh : Surat Pernyataan Artikel Skripsi Tahun 2017,” 2017.

[2] B. Kurniawan, S. Effendi, and O. S. Sitompul, “Klasifikasi Konten Berita Dengan Metode Text Mining,” *J. Dunia Teknol. Inf.*, vol. 1, no. 1, pp. 14–19, 2012.

[3] D. Ariadi and K. Fithriasari, “Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer,” *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.

[4] A. S. Sri Widaningsi, “Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian,” *Semin. Nas. Teknol. Inf. dan Komun. 2018 (SENTIKA 2018)*, vol. 2018, no. Sentika, pp. 23–24, 2018.

[5] G. Ngurah, M. Nata, and P. P. Yudiastra, “Preprocessing Text Mining Pada Email Box Berbahasa Indonesia,” *Konf. Nas. Sist. Inform. 2017*, pp. 479–483, 2017.

[6] V. Amrizal, “Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim),” *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.

[7] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, “Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity,” *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018.

[8] A. H. Setianingrum, D. H. Kalokasari, and I. M. Shofi, “Implementasi Algoritma Multinomial Naive Bayes Classifier,” *J. Tek. Inform.*, vol. 10, no. 2, pp. 109–118, 2018, doi: 10.15408/jti.v10i2.6822.

[9] M. A. Fatmawati, “Klasifikasi Keluhan,” *J. CoreIT*, vol. 3, no. 1, pp. 24–30, 2017, doi: 10.1007/s00766-002-0156-7.

[10] A. Setiyono and H. F. Pardede, “Klasifikasi SMS Spam Menggunakan Support Vector Machine,” *J. PILAR Nusa*

Mandiri, vol. 15, no. 2, pp. 275–280, 2019, doi: 10.33480/pilar.v15i2.693.
[11] U. I. N. Ar-raniry and B. Aceh, “Analisis dan perbandingan kualitas pengelompokan dokumen (document clustering) dengan menggunakan metode k-means dan k-medians bustami yusuf,” vol. 1, no. 2, pp. 1–10, 2015.
[12] V. Amrizal, “Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim),” *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.

ISSN 2085-4552

Digital Image Processing using Texture Features Extraction of Local Seeds in Nekbaun Village with Color Moment, Gray Level Co Occurance Matrix, and k-Nearest Neighbor

Yampi R Kaesmetan[1], Marlinda Vasty Overbeek[2]
[1]Informatic Technique Study Program, STIKOM Uyelindo Kupang, Kupang City, Indonesia
[2]Faculty of Engineering and Informatic, Multimedia Nusantara University, Tangerang, Indonesia
[1]kaesmetanyampi@gmail.com [2]marlinda.vasty@umn.ac.id

Accepted 12 May 2021
Approved 8 August 2021

***Abstract*— The problem in determining the selection of corn seeds for replanting, especially corn in East Nusa Tenggara is still an important issue. Things that affect the quality of corn seeds are damaged seeds, dull seeds, dirty seeds, and broken seeds due to the drying and shelling process, which during the process of shelling corn with a machine, many damaged and broken seeds are found. So far, quality evaluation in the process of classification of the quality of corn seeds is still done manually through visible observations. Manual systems take a long time and produce products of inconsistent quality due to visual limitations, fatigue, and differences in the perceptions of each observer. The selection of local maize seeds in Timor Island, East Nusa Tenggara Province, especially in Nekbaun Village, West Amarasi District with feature extraction with a color moment shows that the mean, standard deviation and skewness features have an average validation of 88% and use the GLCM method which shows the neighbor relationship. Between the two pixels that form a co-occurrence matrix of the image data, namely GLCM, it shows that the features of homogeneity, correlation, contrast and energy have an average validation of 70.93%. The k-Nearest Neighbor (k-NN) algorithm is used in research to classify the image object to be studied. The results of this study were successfully carried out using k-Nearest Neighbor (k-NN) with the euclidean distance and k = 1 with the highest extraction yield of 88% and the results of GLCM feature extraction for homogeneity given 75.5%, correlation of 78.67%, contrast given 65.75% and energy given 63.**

***Index Terms*— corn seeds; color moment; digital image processing; gray level co occurance matrix; k-nearest neighbor**

I. INTRODUCTION

Corn (Zea mays I.) is one of the food crops that contain carbohydrates and is a food that is often consumed by the public in general, in addition to wheat, sweet potatoes and rice. Corn is an agricultural product that is widely planted by farmers because the duration of harvest is faster and plays an important role in the development of the agricultural sector. In Indonesia, especially in East Nusa Tenggara, corn as a staple food is often consumed as a substitute for rice. Where there are sources of protein, carbohydrates, fiber, so that they can help increase endurance, both for consumption of corn that is still not old or that is ready to be harvested. In addition, corn can also be used as food for domesticated livestock such as from young corn stalks, leaves or cobs. Therefore, the demand for corn continues to increase from year to year, this is due to the increase in the economic standard of living of the community and the progress of the animal feed industry so that the quality of corn needs to be considered. There are several factors that affect the quality of corn seeds, one of which is the high level of damage that occurs during the corn shelling process by machine so that many damaged or cracked seeds are found. So far, quality evaluation in the process of classifying corn seed quality is still done manually through direct observation with the naked eye. Visual observation takes a long time so that it can produce products with uneven quality because it has several limitations such as fatigue, and differences in the perception of observers conducting research [1].

Determination of the quality of good corn seed selection has been carried out using supervised learning techniques. One example of a supervised learning method that can be used to determine the quality of corn seeds for seeding is the classification of coffee beans using image processing and fuzzy logic. This research aims to build a digital grading system by using a camera to take pictures of coffee beans samples, then the computer calculates the color and texture. The data test shows that the green color and the entropy, energy and contrast texture characteristics are formed from the co-occurrence matrix with 135 degrees that have been obtained and this is an important parameter of the fuzzy C means in assessing the quality grade of coffee beans [2].

In addition to the classification of coffee beans, there is also research on the determination of the quality of soybeans. Where in the selection of seeds there are broken seeds, pale seeds, dirty and unclean seeds, and broken or cracked seeds due to the drying and firing process that is too long. Soybean quality determination is usually done manually with visual observation. In conducting manual observations can take a long time and produce products with unstable quality because observations are made with the naked eye, fatigue from observers, and different perceptions of each observer when making observations in the field. This research was conducted by using a comparison of image texture extraction using statistical methods of order I (color moment) and statistics of order II (Gray Level Co occurance Matrix - GLCM) for soybean selection. The first order statistic (color moment) shows the probability that the gray level value of the image pixel will appear, while the second order statistic (GLCM) shows the probability of an adjacency relationship between two pixels that form the cohesion matrix of the image data. classification process in determining soybean seeds, and getting an average accuracy of 70.02% [3].

In addition to soybean seed classification research, there is also research on determining the quality of corn seeds for seeding based on color brightness using a support vector machine. The purpose of this study was to classify corn seeds for seeding, so as to find the right corn seeds for seeding based on the brightness level of color. The results of this study were successfully carried out using a support vector machine with a polynomial kernel function and got the highest accuracy, namely yellow corn by 82% and for white corn by 76% [4].

The problem in determining the selection of corn seeds for re-planting, especially corn in East Nusa Tenggara is still an important issue. The declining price of corn in the market caused by damage of corn seeds that will be used for seeds usually occurs when storing corn seeds that have peeled skin making it easier for fungi to grow quickly, especially from the Aspergillus type which has the potential to cause aflatoxins [5]. The selection of corn seeds is often done by farmers by naked eye or manually without looking at the physical characteristics, textures and colors of the corn seeds that will be used for re-seeding. A good selection of corn seeds is corn seeds that are not peeled off, not hollow and black. From the reasons that have been explained, in this research we develop a quality seed selection system using feature extraction based on color moment and GLCM features. While the learning model was developed with supervised learning, k-Nearest Neighbor.

II. Literature Review

A. Corn (Zea mays I.)

Corn is an annual crop, the growth process of corn plants is completed within three months to six months. The first half of the cycle is the vegetative growth stage and the second half is the generative growth stage. Corn is a type of grain food crop (cereal) from the grass family. The following is a systematic (lineage) of corn plants [6] :

Kingdoms : *Plantae* (plant)

Division : Magnoliophyta

Class : Liliopsida

Sub Class : *Commoninidae*

Order : *Poales*

Family : *Poaceaae* (grass tribe)

Genus : *Zea*

Species : Zea mays I.

B. Corn Seeds

Seeds are plant material that will be used for replanting so that they can be used as a means of multiplying similar plants. Corn seeds to be used as seeds must go through a process in such a way that they can be used for the replanting process [6]. The selection of corn seeds must be in accordance with the land to be planted, in order to get good quality corn. Corn seeds must be selected carefully, because corn production depends on seed selection.The criteria for seeds that have low quality include defective or damaged seeds, dull colored seeds, dirty seeds, broken seeds, and small seeds. The following is an explanation table regarding each of the physical quality criteria of corn shown in Table 1 [6].

Table I. Definitions for Corn Seed Physical Quality Criteria

No.	Physical quality	Definition
1	Corn with round seeds	Dried corn kernels that are physically intact without any spots, defects or fungus.
2	Defective seed corn	Corn with defective or damaged seeds due to insect attack or warehouse pests.
3	Corn with broken seeds	Corn with corn kernels that are not intact/damaged due to the threshing or shelling process.
4	Dull seed corn	Colored corn kernels tend to be dirty or dark

C. Pre-Process

This stage is carried out to obtain data accuracy from the image of corn seeds that will be sampled, this process is to prevent data inaccuracies to get actual data. RGB image or commonly called true color is an image that can represent the color of an object that resembles the original by combining three colors that are often used, namely red (R), green (G) and blue (B). Each pixel of an RGB image has three channels that represent each component of the basic color [7]. A gray image is a digital image that has only one channel value for each pixel, in other words, the value of the red = green = blue part. This value is used to indicate the level of intensity. The color of the three grayscales is a gray color with various levels from black to white. Grayscale images can be obtained from RGB images. The intensity value of the grayscale image is calculated from the intensity value of the RGB image using the eq. 1.

$$Gray\ Value = R + G + B / (\alpha + +) \qquad (1)$$

With the value of = the value for R (0.35), β= the value for G (0.25), = the value for R (0.4), so the value of + δ + = 1. The image that has been changed to The gray scale will be processed to remove noise using a median filter by finding the average of the image pixel values that have been sorted by the eq.2 :

$$f\,(y,x) = median(p,q)\epsilon syx\,(g(p,q)) \qquad (2)$$

Where *f(y,x)* = weight of result at position *(y,x)*, *g(p,q)* = element of gauss kernel matrix at position *(p,q)*.

D. Digital Image

Digital images can be interpreted as light intensity on two sides and can be expressed in two dimensions *f(x,y)* where is the light intensity in discrete form about the axis *x* nor *y* which is the position of the coordinate point while *f* is the amplitude at position *(x,y)* which is often known as intensity or grayscale [8]. In assigning a value of a discrete intensity from *0* to *255*, as well as the values of *x,y* and *f(x,y)* remain in a certain range or area but are in limited quantities. The image taken from the camera and the process of limiting the input of a wide set in discrete form is called a digital image. A digital image is composed of a number of gray level values called pixels at a certain position.

In the translation of light intensity can be calculated by the equation of two dimensions *f(x,y)* is:

$$0\ f(x,y) < \infty \qquad (3)$$

For example, f is a 2-dimensional digital image measuring *NxM*. So that it can be spread *f* in a matrix can be seen in the figure below, where *f(0,0)* is in the upper left corner of the matrix, while *f(n-1, m-1)* is in the lower right corner.

$$f(x,y)\begin{bmatrix} f(0,0) & f(0,1) & f(0,M-1) \\ - & - & - \\ f(N-1,0) & f(N-1,1) & f(N-1,M-1) \end{bmatrix} \qquad (4)$$

E. Texture Analysis

The textured analysis used is a form of the intrinsic characteristics of an image form and is closely related to the level of roughness, granulation, and regularity of the structural arrangement of pixels. The textural aspects of an image can be used as the basis for segmentation, classification, and image interpretation [9].

So that the image texture can be interpreted as a function of the spatial variation of pixel intensity (gray value) in the image. Based on their shape, textures can be classified into two classification:

a. Macrostructure

The shape of the macrostructure has a local pattern repeating periodically in an image area, usually found in man-made patterns and tends to be easy to represent mathematically.

b. Microstructure

In microstructural texture, local and repeating patterns do not occur so clearly, so it is not easy to provide a comprehensive definition of texture.

Fig. 1 is an example of a texture that shows the difference between macrostructure and microstructure textures.

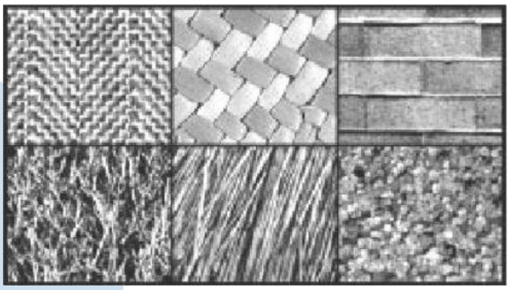


Fig. 1. Examples of visual textures [9]

F. Color Moment

Color moment is a form of representation of taking features based on the characteristics of the color of an image. From a histogram, it can show the probability of the occurrence of the gray level value of pixels in an image [9]. Color moments have 3 characteristics, there is:

1). Mean (μ)

Mean(μ) describes the shape of the size of the dispersion of an image. The form of the equation to calculate the Mean (μ) is:

$$\mu = \Sigma_n f_n p(f_n) \qquad (5)$$

Where (μ) is the average value of a color in the image, n is the total number of pixels in the image and f_n is a value of gray intensity, while p(f_n) shows the histogram value (probability of the occurrence of that intensity in the image).

In developing applications, the authors use the Firebase platform. Firebase is a service from Google that is used to facilitate application developers in developing applications. One of the functionalities of Firebase is a realtime database service that we use to store allergen data. Realtime database has the ability to store and synchronize application data in milliseconds.

The database will be hosted in the cloud. Data will be saved as JSON and then synchronized in realtime to each client that has been connected so that all clients can receive the latest data updates automatically.

In this study, the database stores allergen data in the form of id, composition, and general name of the allergen. Fig. 5 is an example of a storage format in the database. "Komposisi" is the name of the allergen (processed / technical) which may appear on food packaging, "NamaGeneral" is the name of the main allergen of the ingredient, for example in Fig. 5 ingredients is "lactalbumin" which is processed from milk (susu) so that it has a generalized name "susu".

2). Standard deviation or standard deviation

The standard deviation or standard deviation is the most frequently used measure of variation (variation) of statistical data. The standard deviation or standard deviation is the square root of the variance. Form of calculation in standard deviation or standard deviation (σ) that is:

$$\sigma = \sqrt{\frac{1}{n}\sum_{n}^{j=1}(f_n - \mu)^2 p(f_n)} \qquad (6)$$

3). Skewness (α_3)

Skewness (α_3) shows the degree of skewness (a measure of the degree of asymmetry) relative to the histogram curve of an image. Solution to calculate Skewness (α_3) that is:

$$\alpha_3 = \frac{1}{\sigma^3}\sum_n (f_n - \mu)^3 p(f_n) \qquad (7)$$

G. Gray Level Co-occurance Matrix

In analyzing a pixel to obtain statistical characteristics by solving the probability value or often called the probability of an image which can be obtained from the neighboring relationship between two pixels at a certain distance and angle orientation. By using GLCM, the calculation process can work by forming a co-occurrence matrix from image data, then it will be continued by determining the characteristics as a function of the matrix [9].

The number of occurrences of joint calculations or co-occurrence, where it is the number of occurrences of one level of neighboring pixel values with one level of other pixel values within a certain distance (*d*) and angle orientation (θ). Distance is expressed in pixels and orientation is expressed in degrees. Orientation is formed in four angular directions with angular interval 45°, that is 0°, 45°, 90°, 135°. While the distance between pixels is usually set at 1 pixel.

In this case, the co-occurrence matrix is a square matrix with the number of elements as much as the square of the number of pixel intensity levels in the image. Each value of the point *(p,q)* in the oriented co-occurrence matrix contains the probability of occurrence of a value of *p* neighbors with a pixel worth to *q* at distance *d* and the orientation of and (*180-*). The following is an illustration of Order II Statistical Feature Extraction which is shown in Fig 2.

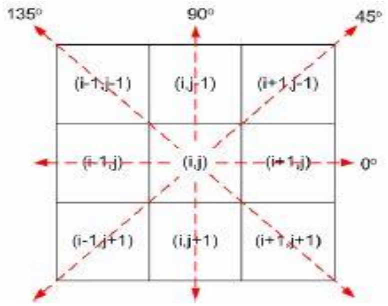


Fig. 2. Illustration of GLCM Feature Extraction [9]

In the form of completion of the GLCM feature extraction, and the co-occurrence matrix is obtained from each angle, the average co-occurrence matrix (Mavg) will be shown in the equation 8:

$$M\ avg = (M(0) + M(45) + M(90) + M(135)) / 4 \qquad (8)$$

GLCM (Gray Level Co-occurance Matrix) has 14 haralick characteristics or features, but what is taken in this research are 4 characteristics, namely:

1). Angular second moment (homogeneity)

Angular second moment(homogeneity) shows the homogeneity of the image. The form of the equation shown in equation 9:

$$ASM = \sum_i \sum_j \{p(i,j)\}^2 \qquad (9)$$

Where *p(i,j)* represents the value on the row *i* and column *j* on the co-occurrence matrix.

2). Contrast

Where in the form of contrast features can show the size of the spread (moment of inertia) of the image matrix elements. If it is located far from the main diagonal, the contrast value is large. Visually, the contrast value is a measure of the variation between degrees of gray in an image area. The form of the equation shown in equation 10:

$$CON = \sum_k k^2\left[\sum_i \sum_j p(i,j)\right] \qquad (10)$$
$$|i - j| = k$$

3). Correlation

Where in Correlation can be shown with a linear dependence of the degree of gray in the image that has been obtained and can provide clues to the existence of linear structures in the image. Scoreμ shows the average color in the image shows the square root of the variance (standard deviation). The form of the equation shown in equation 11:

$$COR = \frac{\sum_i \sum_j (ij) - \mu_x \mu_y}{\sigma_x \sigma_y} \qquad (11)$$

4). Energy

Energy show uniformity. Energy will be high when the pixel values are similar to each other otherwise it will be small indicating the value of the normalized GLCM is heterogeneous. The maximum value of energy is 1, which means that the distribution of pixels is in a constant condition or in a periodic (not random) shape. The value of d indicates the distance between two pixels. The form of the equation shown in equation 12:

$$EN = \sum_i \sum_j p^2 d\,(i, j) \qquad (12)$$

H. K-Nearest Neighbor

In understanding algorithm *k-Nearest Neighbor* (k-NN) which is a method for classifying objects based on the learning data that is closest to the object and already has a class label. Where data in learning is projected into a multidimensional space, where each dimension represents a feature of the data. This space is divided into sections based on the classification of learning data. A point in this space is marked class c if class *c* is the most common classification found in *k* the nearest neighbor of the point. Near or far neighbors are calculated based on Euclidean distance.

When compared with other classification methods, this method has a fairly high level of accuracy because the incoming data will be classified based on the similarity of existing characteristics from the previously classified data. However, on the algorithm *k*-NN needs to determine the value of the parameter *k* (number of nearest neighbors) and distance-based learning it is not clear what type of distance to use and which attributes to use to get the best results [10]. The *k*-NN for Euclidean distance is as count as follows:

$$D_{st} = \sqrt{\sum_{j=1}^{n} (X_{sj} - Y_{tj})^2} \qquad (13)$$

With $d(x,y)$ is a distance between points in training data data X_i and testing data points Y_i which will be classified, where $x=x_1,x_2,\ldots,x_i$ and $y=y_1,y_2,\ldots,y_i$ and i represents the attribute value as well as n is an attribute dimension.

III. Result and Discussion

In this research, we used color moment and GLCM as a feature extraction and k-NN as a model classifying. The moment features used are: mean, standard deviation and skewness, while the haralick (GLCM) features used are: homogeneity, correlation, contrast, energy. The analysis of the results of the classification test with the K-Nearest Neighbor in terms of the use of the k value and the distance measurement method (Euclidean), as well as the analysis of system computational time. This system was developed with Matlab 10th student version.

The image of corn seeds taken in this study was 100 corn seed image data, with initial image dimensions of 2484 x 2134 pixels and saved in JPG file format. The image that will be sampled will be resized to 300 x 300 pixels, so that it gets an area of interest from the image, this image data will be called into Matlab to get the matrix data to be processed. The following are the characteristics of the image of decent and unworthy corn seeds to serve as seeds. Table 2 shown the explanation of it.

Table II. Decent And Unworthy Corn Seed

No.	Decent seeds	Unworthy seeds
1.	Clean and whole seeds	Dirty seeds
2.	Seeds that are not black	Black seeds
3.	The unbroken seed	Broken seeds
4.	Undamaged seeds	Broken seeds
5.	Seeds that don't have holes	Hollow seeds

A. Feature extraction

In this process, the image that has been preprocessed will be extracted using *color* moments and GLCM.

B. Color moment

Color moment shows the probability of occurrence of a pixel's gray degree value in an image. The previously 100 x 100 image matrix will be converted into a 1 x 3 vector with a feature extraction process using HSV images. At this stage the image will also calculate the value of each feature. The characteristics used in this extraction include mean, standard deviation and skewness. The amount of data used is 100 corn image data so that the resulting 100 vectors. This vector is the input dataset for the classification process. The following is the syntax or command line in Matlab to get the result of color moment.

```
b=imresize(image1,[300
300]);
gbrHSV = rgb2hsv(b);
c1                          =
imresize(gbrHSV,[300
300])
```

C. GLCM

GLCM shows the probability of adjacency relationship between two pixels at a certain distance and orientation angle. The previously 100 x 100 image matrix will be converted into a 1 x 1 vector with this extraction process using a grayscale image. At this stage the image will also calculate the value of each feature and also each direction. The characteristics used in this extraction include homogeneity, correlation, contrast and energy, while the directions used are 0^0, 45^0, 90^0, 135^0. The amount of data used is 100 corn image data so that the resulting 100 vectors. This vector is the input dataset for the classification process. Following is the syntax or command line in Matlab to get the results from GLCM.

```
b=imresize(image1,[300
300]);
gbrgray=rgb2gray(b);
c1=imresize(gbrgray,[300
300
```

D. Classification using k-NN

The classification using k-NN is divided into two processes, namely the training and testing process. The training process is used to produce a classification model with k-NN which will later be used as a reference for classifying the quality of corn seeds to be used as seeds with new raw data. In this study, the authors used the Euclidean distance with k values of 1, 3 and 5.

The data used in this research is Lamuru corn image data typical of Nekbaun Village, West Amarasi District with 100 seed image data for color moment feature extraction and GLCM. The image data is then divided into 75 samples as training data and 25 samples as test data stored in an array. Each image data has dimensions of 1 x 3 for feature extraction of order I (color moment) of each feature (mean, standard deviation and skewness) and dimensions of 1 x 1 for extraction of order II features (GLCM) of each feature (homogeneity). , correlation, contrast and energy). Training and testing data through preprocessing and feature extraction stages. The result of feature extraction of each image data has dimensions of 1 x 3 (color moment) and 1 x 1 (GLCM). All data are then arranged based on the composition of the training for k-NN based on the Euclidean distance with k values of 1, 3 and 5. The training data has a vector data dimension of a matrix of 100 x 3 corn seed image for color moment and the second training has a matrix vector data dimension of 100 x 1 corn seed image for GLCM. The target of the results of this study is to find the best results on k-NN based on the Euclidean distance with k values of 1, 3 and 5 nearest neighbors, which will be used to identify each test data, whether it is in class 0 or 1(unworthy and decent). The scenario model used is 4- fold cross validation. The following is a scenario model in calculating the dataset used in determining the quality of corn seeds shown in Table 3.

TABLE III. MODEL SCENARIOS

Scenario name	*Dataset* (100 data)				Explanation
Scenario I	data 1-25	data 26-50	data 51-75	data 76-100	data 1-25 as test data, data 26-100 as training data.
Scenario II	data 1-25	data 26-50	data 51-75	data 76-100	26-50 data as test data, data 1-25 and data 51-100 as training data.
Scenario III	data 1-25	data 26-50	data 51-75	data 76-100	data 51-75 as test data, data 1-50 data 76-100 as training data.
Scenario IV	data 1-25	data 26-50	data 51-75	data 76-100	data 76-100 as test data, data 1-75 as training data.

E. System Interface

The interface is a media liaison between the system and the user. System operation will start on this system interface page, making it easier for users to use this application. The following system interface can be seen in Figure 3.

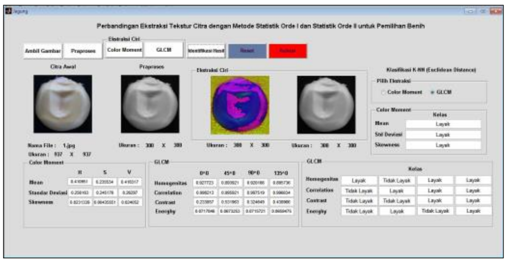


Fig. 3. System interface

F. Test Result

System accuracy testing is done by measuring the performance of the corn seed quality determination system based on texture. In this research, the image tested is the image of corn. This test is carried out using the k-NN classification (euclidean distance) with k values of 1, 3 and 5. The test is seen from how well the classifier predicts the quality of corn seeds that are decent or unworthy for a seeds. Testing is done by looking for the value of sensitivity, specificity and accuracy of the classification system to determine the accuracy of this system. Based on the results of system testing that has been carried out, the sensitivity, specificity and accuracy values of the system are obtained. The result of this image processing is the generation of numerical data from each image of the corn seeds which will be separated into decent and unworthy corn seeds. The results of the average color moment feature extraction processing can be seen in Table 4.

TABLE IV. AVERAGE RESULT USING COLOR MOMENT FEATRURE EXTRACTION

Characteristic	The average value of suitable corn seeds for planting	The average value of corn seeds is not suitable for planting
mean	0.3516	0.3347
Standard deviation	0.2347	0.2189
Skewness	0.0174	0.0126

In this research, the k-NN classification is used with the distance used is Euclidean distance and uses feature extraction of order I or color moment. The data scenario used is a 4-fold cross validation. The accuracy results obtained from the k-NN classification for the mean is 90%, the standard deviation is 88% and the skewness is 86%. Validation is repeated once for each feature and the average is taken. From the validation data, it was found that image extraction to distinguish viable and unfit corn seeds was feasible to use. For more details, it can be seen in the differences in sensitivity and specificity tests for each feature shown in Figure 4.

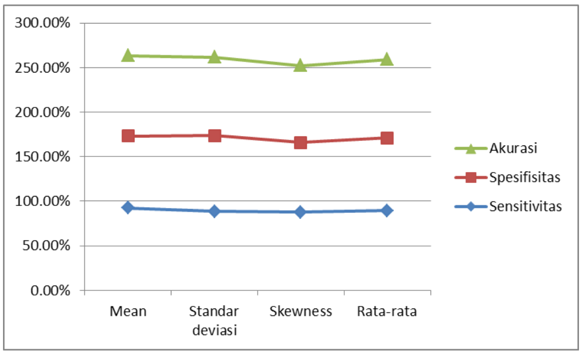


Fig. 4. Average result with color moment

From Figure 4, it can be seen that the mean characteristic given sensitivity is 92.42%, specificity is 80.83%, accuracy is 90%; the standard deviation characteristic given sensitivity is 88.67%, specificity is 85%, accuracy is 88% and the skewness characteristic given sensitivity is 87.76%, specificity of 78.33%, accuracy of 86%. Of the 3 characteristics, the average value for sensitivity is 89.61%, the average value for specificity is 81.38% and the average value for accuracy is 88% so it can be said that the classification technique with k-NN has given good results in classifying images with textures using the color moment feature.

For GLCM feature extraction, given different result if the model classification done with k-NN algorithm. The average result, shown in figure 5.

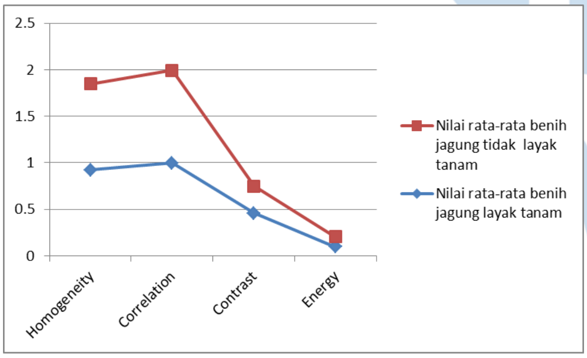


Fig. 5. Average result with GLCM

From Figure 5 it can be seen that the highest average texture feature is correlation because it has an average value of decent corn seeds is 0.9967 and unworthy corn seeds is 0.9969, this means that the size of the linear dependence of the gray level of the image gives an indication of the existence of linear structures in the very high image. While the feature with a low average value is energy because it has an average value of decent corn seeds is 0.0982 and uunworthy corn seeds is 0.1056 this means the uniformity on the texture of the image is less.

The k-NN classification is used with the distance used is the Euclidean distance and uses feature extraction of order II or GLCM. The validation results obtained from the k-NN classification for each feature are 63.82% for the energy feature, 75.5% for the homogeneity feature, 65.75% for the contrast feature, and 78.67% for the correlation feature. Validation is repeated once for each feature and the average is taken. From the validation data, it was found that image extraction to decent and unworthyt corn seeds was feasible to use. For more details, it can be seen in the differences in sensitivity and specificity tests for each feature shown in Tables 5 to 8.

TABLE V. SENSITIVITY AND SPECIFICITY OF HOMOGENEITY FEATURE

Direction	Sensitivity	Specificity	Accuracy
0º	100%	100%	100%
45º	32.10%	70.83%	51%
90º	96.25%	19.17%	81%
135º	96.25%	19.17%	81%
Average	81.15%	52.29%	75.5%

TABLE VI. SENSITIVITY AND SPECIFICITY OF CORRELATION FEATURE

Direction	Sensitivity	Specificity	Accuracy
0º	77.5%	40%	48%
45º	100%	100%	100%
90º	100%	17.5%	83.67%
135º	100%	17.5%	83%
Average	85%	53.12%	78.67%

TABLE VII. SENSITIVITY AND SPECIFICITY OF CONTRAST FEATURE

Direction	Sensitivity	Specificity	Accuracy
0º	41.10%	80.41 %	49%
45º	27.37%	95.83%	42%
90º	100%	100%	100%
135º	98.80%	44.58 %	86%
Average	61.50%	81.04%	65.75%

TABLE VIII. SENSITIVITY AND SPECIFICITY OF ENERGY FEATURE

Direction	Sensitivity	Specificity	Accuracy
0º	41.10%	80.41 %	49%
45º	95.23%	40%	84%
90º	45.59%	76.35%	52.29%
135º	63.69%	95%	70%
Average	61.40%	72.91%	63.82%

From Table 5 to Table 8, it can be seen that the homogeneity feature for sensitivity is 81.15%, for specificity is 52.29%, and accuracy is 75.5%; for correlation feature, for sensitivity is 85%, for specificity is 53.12%, and accuracy is 78.67%; contrast feature that is for sensitivity is 61.50%, for specificity is 81.04%, and accuracy is 65.75%; and energy feature, for sensitivity is 61.40%, for specificity is 72.19%, and accuracy is 63.82%. So from the 4 features in GLCM, it can be said that the classification technique with k-NN has given good results in classifying images with textures using haralick features. The following is a table of averages of sensitivity and specificity tests for GLCM corn seed images, which can be seen in Table 9.

TABLE IX. AVERAGE OF SENSITIVITY AND SPECIFICITY USING GLCM FEATURES

Haralick features	Sensitivity	Specificity	Accuracy
Homogeneity	81.15%	52.29%	75.5%
Correlation	85%	53.12%	78.67%
Contrast	61.50%	81.04%	65.75%
Energy	61.40%	72.91%	63.82%
Average	72.27%	64.84%	70.93%

From table 9 it can be seen that the average sensitivity, specificity and accuracy for GLCM testing is 72.27% for sensitivity, 64.84% for specificity and 70.93% for accuracy.

From the two classification test results in Table 4 and Table 9, the extraction is used first-order statistical features (color moment) and extraction using second order statistical features (GLCM) using the k-NN model classification with euclidean distance for the selection of corn seeds given a good results. Because with color moment feature extraction, given 88% of accuracy and 70.93% accuracy if using GLCM features extraction.

IV. CONCLUSION

In this research for the selection of local corn seeds in Nekbaun Village, Amarasi Barat District, East Nusa Tenggara with Color Moment feature extraction, it showed that the mean, standard deviation and skewness characteristics had an average validation of 88% and GLCM feature extraction showed that the homogeneity characteristics, correlation, contrast and energy have an average validation of 70.93%. From these results, it can be concluded that feature extraction with Color Moment is better than GLCM feature extraction in classifying images into decent seed class and unworthy seed class.

REFERENCES

[1] Supriatna, A. 2009. Identification of Physical Quality of Corn using Digital Image Processing Technology and Artificial Neural Networks. 31st volume. Bogor (ID): Bogor Agricultural Research and Development Agency.

[2] Maria, 2010. Classification of Coffee Bean Quality using Image Processing and Fuzzy Logic. Thesis. Surabaya (ID): Ten November Institute of Technology Surabaya.

[3] Kaesmetan Y R," Comparison of Image Texture Extraction for Selection of Soybean Seeds with Order I and Order II Statistical Methods", Journal of Information Technology Hoaq 2019, pp.996-1006, December,2019.

[4] Bilaut, CA. 2015. Determining the quality of corn seeds for seeding based on the brightness level of the color using a support vector machine. Thesis. Kupang (ID): Stikom Uyelindo Kupang.

[5] Agrology, 2012. Physiological Quality of Corn to Seedlings, Journal Volume 11th. Bogor (ID): IPB Bogor.

[6] Anggraini, 2008. Estimating the Physical Quality of Corn with Artificial Neural Networks using Color Parameters in Digital Image Processing. Journal Volume 1. Padang (ID): UNAND Padang.

[7] Gonzalez, CR. Woods, R. 2007. Digital Image Processing. Prentice Hall Publishers. New Jersey (US): USA

[8] Gonzalez, CR. Woods, R. 2007. Digital Image Processing. Prentice Hall Publishers. New Jersey (US): USA

[9] Ardi, 2011. Texture Analysis using the GLCM (Gray Level Co-occurance Matrix) method and feature extraction. Bandung (ID): ITT Bandung

[10] Zhu, S. Dong, W. & Liu, W. 2015. Hierarchical Reinforcement Learning Based on KNN Classification Algorithms. International Journal of Hybrid Information Technology. Journal Volume 8. Pages 175-184.

ISSN 2085-4552

Algorithm with SVD The Implementation of the Weight Product (WP) Method on the Best Employee Selection

The Decision Support System of the Best Employee Selection Implemented by Weight Product Method at PT. Autogrill Services Indonesia

Komang Redy Winatha[1], I Nyoman Tri Anindia Putra[2], Naufal Akbar Ihsan Baedlawi[3]
Informatics Engineering, STMIK STIKOM Indonesia, Denpasar, Bali - Indonesia
[1]redywin@stiki-indonesia.ac.id, [2]trianindiaputra@stiki-indonesia.ac.id, [3]naufalbaedlawi@gmail.com

Accepted 3 June 2021
Approved 16 June 2021

***Abstract*— PT. Autogrill Services Indonesia is a private company engaged in the food and beverages selling. There are 8 outlets and 379 employees. To achieve maximum performance within the company environment, PT Autogrill Services Indonesia gives an appreciation to employees in the form of the best rewards every month and year calculated based on certain criteria. PT AutoGrill Services Indonesia needs to have a decision support system to simplify the decision-making process. To meet these needs, a web-based decision support system for selecting the best employees was designed using the weight product (WP) method at PT Autogrill Services Indonesia. The design stage includes needs analysis, context diagrams, data flow diagrams, and designing database tables. This system is web-based, using the programming language PHP and MySQL as database storage. The main features contained in this system include processing user data, outlets, employees, criteria, periods, alternatives, scores, and the calculation of monthly and annual winners. Based on the test results, all system functionality components can run well and by expectations.**

***Index Terms*— DSS, Decision, Reward, Weight Product**

I. INTRODUCTION

Employees are one of the main assets for a company. The existence of employees plays an important role either in the implementation or operations in order to achieve business targets[1]. Therefore, companies should always maintain and improve employee performance by means of training, establishing the solid work team, career paths and fulfillment of employee rights and provide rewards or awards for superior employees. Reward is one of the strategies ran by organization or company to motivate its human resources to make more contributions to the company [2]. This is reinforced by the statement by Mangkunegara [3] formulates that performance belongs to inseparable two things namely ability and motivation. The point is in order to achieve optimal performance quality, it takes not only about the ability, but also employee's motivation. Why? Because, if the company has quality resources, it can directly reach the predetermined target. Giving rewards for employees can be a very important factor because it can motivate employees to work harder and improve performance [4]. So far, most companies have given rewards to superior employees to maintain the best performance of employees, including those carried out by foreign private companies namely PT. Autogrill Services Indonesia.

PT. Autogrill Services Indonesia is a private company engaged in the hospitality sector and food and beverages selling, which is a branch of HMSHost International, headquartered in the Netherlands. PT. Autogrill Services Indonesia located at Ngurah Rai International Airport and has 379 employees as of 23 January 2020 consisting of 50 employees of back office management and 8 outlets/restaurants, which divided into: 37 employees of The Coffee Club Departure, 31 employees of Two Dragons, 58 employees of Two Tigers, 25 employees of La Place Landside, 97 employees of La Place Airside, 60 employees of House of Beans, 6 employees of Urban Food Market, and 15 employees of The Coffee Club Arrival. In order to give appreciation to superior employees, each outlet every month gives awards to the best employees (employee of the month). Winners from each monthly outlet will then be contested again to determine the best employee of the year.

The results of interviews and observations with managers and outlet leaders obtained information that the determination of the best employees so far is still subjective based on the manager's personal assessment and then stored them all in the manager's notebook. The limited capabilities possessed by human resources in this case pointed out on decision making, it became a problem when determining who will be the best employees because there are criteria and alternatives that must be counted. Then, there is no system which expected to manage criteria data and calculations in order to produce the best employee information, so that the employees who do not fulfil the requirement can possibly win to be the best employee and might raise the high intense of toxic work culture among the employees, instead of having difficulty of analysing the employees performance data review within year to year.

Based on the explanation above which have been described previously, a computerized decision support system (DSS) is needed in order to determine the best monthly and annual employees according to predetermined criteria and requirement. This system can be a tool to assist the companies in making decisions. DSS is a form of computer-based information system specifically developed to support problem solving [5]. DSS can also be interpreted as a computerized information system provides interactive support for business people during the decision-making process [6]. Another definition explains that DSS is a computer-based system that utilizes certain data and models to help on decision making and problem solving [7]. The aim of The DSS developing is not intended for the final and absolute decision making, but is built to evaluate an opportunity that might be taken through analysis using the available methods. [8].

Weight product is a method used and implemented on a computerized decision support system (DSS). It is a decision-making method with multi-criteria which used to solve the problem [9]. The selection of this method is based on being more efficient, because the time required for the calculation is relatively shorter [10][11] compared to other methods. This method is able to select alternatives and it is also has the advantages in weighting techniques so that it can be called the easiest method to design among other methods [12].

In addition, recent studies have also examined the effectiveness of the weight product (WP) method compared to other methods. Among them are research by Setyawan, Arini & Akhlis [13], which analyses the comparison of the WP and SAW methods in supporting the decision of the new employee recruitment process. The results showed that the WP method gave more accurate results than the SAW method. The same method was also examined by Purnomo & Rozi [14] with a different case study, namely the selection of the best graduates. This study uses two test indicators, namely testing based on the speed of access time and testing RSD (Relative Standard Deviation). The results of the test on the speed of access time and RSD testing, the WP method is more recommended because it is more optimal.

Based on the explanation of the problems and support from related studies, a research was conduted entitled "Implementation of the Weight Product (WP) Method in the Selection of the Best Employees". The purpose of this research is to implement the WP method in a web-based computerized system that can help companies, in this case, PT. Autogrill Services Indonesia in selecting the best employees. The implementation of the weight product (WP) method is considered very relevant in case studies of selecting the best employees based on recent studies that also use the same method.

II. METHOD

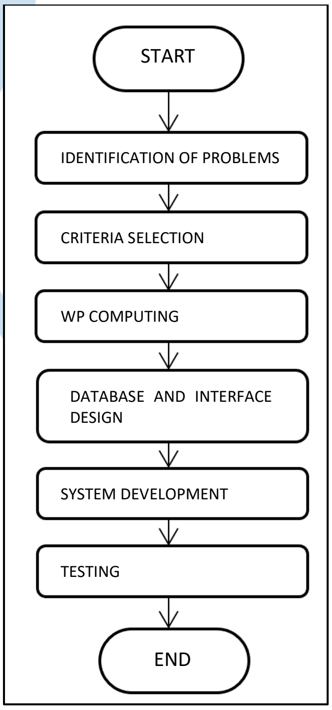
The method used in this research consists of several stages, namely:

A. *Data Collection*

Data collection was carried out in two techniques, namely primary data and secondary data. Primary data were obtained through direct observation and interviews with the outlet manager of PT. Autogrill Services Indonesia. While secondary data were obtained through documentation and literature from various media such as: internet, journals, and those related literature books.

B. *Research Stage*

The research steps begin with problem identification, criteria selection, weight product computation, database and interface design, system development, and testing (Figure 1).



Fugure 1. Research Stage

C. *Determining the Problem and Selection of Criteria*

This research departs from the problem of constrained processes in selecting the best employees. The assessment criteria and the weight of each criterion are shown in Table 1.

TABLE 1. CRITERIA AND WEIGHT ASSESSMENT

No.	Kriteria	Nilai Bobot
1	Neatness	2
2	Skills	3
3	Absence	4
4	Lateness	4
5	*Complain*	5

D. *Weight Product (WP) Computing*

WP computing is carried out through the following stages:

1) Deteremine criteria
The criteria used as a reference in decision making and the nature of the criteria can be seen in Table 2.

TABLE 2. CRITERIA AND ITS TRAIT

Criteria	Trait
C1 : Neatness Neatness standards are needed as a characteristic of a company to foster consumer confidence in the company.	Cost Reason: judging from the quantity of attributes used or not according to company rules.
C2 : Skill Employee skills are very important in completing various jobs, besides that employee skills also affect the business process activities that are being undertaken by a company.	Cost Reason: judged by the quantity of damage or loss of work equipment.
C3 : Absence The absence of employees is a disadvantage for the company, because the absence of one or more workers will affect the performance of the team in general.	Cost Reason: the more employees do not come to work, the lower the performance appraisal.
C4 : Delay Employee tardiness is an disciplinary action that results in disruption of activities when changing work shifts between company employees. This can cause a loss to your teammates.	Cost Reason: the more often employees are late for work, the lower the assessment score will be.
C5 : Complaint Customer complaints against employees are detrimental for the company. Because complaints are things which show that the employee cannot work properly according to service standards that exist in the company.	Cost Reason: the more complaints customers give to employees, the lower the employee's performance appraisal.

2) Weight Normalization
Normalization of weights is calculated by dividing the weight of each criterion by the sum of all the weights of the criteria. Normalization of weights using the equation:

$$Wj = \frac{W}{\Sigma Wj} \quad (1)$$

3) Determining the value of the vector S
The value of the vector S, is calculated using the following equation:

$$Si = \Pi_{j=1}^{n} X_{ij}^{Wj} \quad (2)$$

4) Determining the value of vektor V
The vector value to be used calculates the preference (Vi) for ranking. Here is the equation:

$$Vi = \frac{\Pi_{j=1}^{n} Xij^{wj}}{\Pi_{j=1}^{n} Xj^{wj}} \quad (3)$$

5) Ranking the vector values V
The next step is to rank the value of Vi from each alternative. The alternative with the highest Vi value is the best alternative in supporting the decision.

E. *Database and Interface Design*

In designing the table structure, this decision support system consists of: tb_alternatif, tb_pemenang, tb_vector, tb_karyawan, tb_outlet, tb_normalisasi_bobot, tb_perhitungan, tb_kriteria, tb_user, tb_temp_skor, tb_skor, dan. The relationship between tables is described as follows:

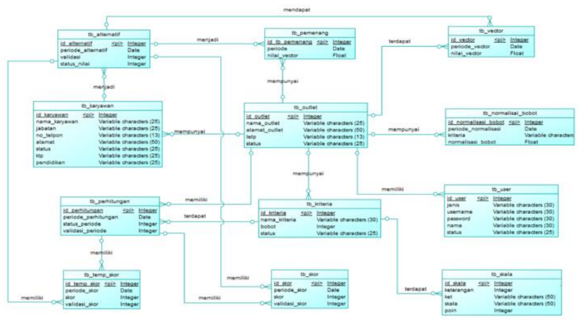


Figure 2. *Conceptual Data Model*

The system interface consists of the main page, login page, profile page, outlet manager page, alternative data page, criteria data page, calculation process page, monthly winner list page, annual winner list page, outlet data page and user data page. The designed interface will display information that can be accessed by outlet managers, operational managers and employees of PT. Autogrill Services Indonesia.

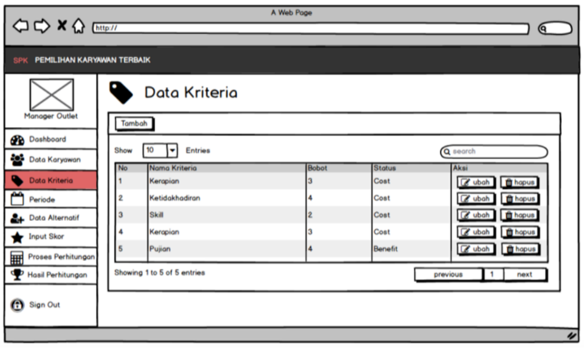


Figure 3. Interface Design

F. *Testing*

Testing this decision support system using the black box testing method. This method is used to evaluate software without having to pay attention to details and only checks the output value based on the input value [17]. Blackbox testing is carried out to find several errors, namely: (1) there are missing or even incorrect functions; (2) interface design errors; (3) errors in the data structure or database; (4) faults in terms of performance; and (5) termination and initialization errors [18].

III. RESULTS AND DISCUSSION

The initial stage in the study was observation and interviews. Observations were made by coming directly to PT. Autogrill Services Indonesia to observe and find out further the information about problems and solutions to these problems. The next data collection is to conduct an interview with Ahyat Yatendra as the outlet manager of PT. Autogrill Services Indonesia. From all data collection techniques, information is obtained about the best prospective employee data, and what criteria are needed in determining the best prospective employee.

To determine whether or not an employee is eligible to win the best employee competition at PT. Autogrill Services Indonesia then determined 5 criteria that must be met by employees, which later these criteria will be used as a reference for the company in making decisions.

TABLE 3. CRITERIA

No	Criteria
1	Neatness
2	Skills
3	Absence
4	Lateness
5	*Complaint*

From the provisions of these criteria, then the weight of each criterion is determined which determines the level of importance.

TABLE 4. WEIGHT CRITERIA

No	Criteria	Weight Value	Status
1	Neatness	2	*Cost*
2	Skills	3	*Cost*
3	Absence	4	*Cost*
4	Lateness	4	*Cost*
5	*Complaint*	5	*Cost*

After determining the criteria and the weights of these criteria, the next step is to describe the definition and weighting of the rating scale of these criteria.

A. *Neatness*

Neatness is needed as a characteristic of a company to foster consumer confidence in the company.

TABLE 5. NINETY ASSESSMENT SCALE

Scale	Description
1	Not wearing attributes according to company rules more than 8 times
2	Not wearing attributes according to company rules as much 6 – 8 times
3	Not wearing attributes according to company rules as much 3 – 5 times
4	Not wearing attributes according to company rules as much 1– 2 times
5	Always wear attributes according to the rules

B. *Skills*

Employee skills are an important indicator in completing various jobs, besides that employee skills also affect the business process activities that are being undertaken by a company.

TABLE 6. SKILL ASSESSMENT SCALE

Scale	Description
1	Damage or lose equipment more than 8 times
2	Damage or lose equipment 6 -8 times
3	Damage or lose equipment 3 -5 times
4	Damage or lose equipment 1-2 times
5	Never damage equipment or remove it altogether.

C. *Absence*

Employee absence is a disadvantage for the company because the absence of one or more workers will affect team performance.

TABLE 7. ATTENDANCE ASSESSMENT SCALE

Scale	Description
1	Absent more than 8 times
2	Absent 6-8 times
3	Absent 3-5 times
4	Absent 1-5 times
5	Never absent at all

D. Lateness

Delays can result in disruption of activities when changing work shifts between company employees.

TABLE 8. LATE ASSESSMENT SCALE

Scale	Description
1	Late more than 8 times
2	Late 6 - 8 times
3	Late 3 – 5 times
4	Late 1 – 2 times
5	Never late at all

E. Complaint

Complaints are things that show that the employee cannot work properly according to service standards that exist in the company.

TABLE 9. COMPLAIN RATING SCALE

Scale	Description
1	More than 8 complaints from customers
2	Complaints 6-8 times from customers
3	Complaints 3-5 times from customers
4	Complaints 1-2 times from customers
5	Never got any complaints from customers

Calculation

Examples of data from employees who will be selected to be the best employees are with 3 employees as follows:

TABLE 10. EMPLOYEE DATA

Criteria	Naufal	Dwi	Wawan
Neatness	1	2	3
Skills	2	1	4
Absence	1	2	3
Lateness	1	2	2
Complaint	2	3	1

TABLE 11. EMPLOYEE VALUE DATA WITH SCALE

Criteria	Naufal	Dwi	Wawan
Neatness	2	2	3
Skills	2	2	3
Absence	2	2	3
Lateness	2	2	2
Complaint	2	3	2

$$\text{Score} = \sum(\text{criteria weight } x \text{ rating scale value})$$

TABLE 12. EMPLOYEE FINAL VALUE DATA

Criteria	Naufal	Dwi	Wawan
Neatness	4	4	6
Skills	6	6	9
Absence	8	8	12
Lateness	8	8	8
Complaint	10	15	10

The steps for calculating the Weight Product (WP) method from the above case examples are as follows:

1) Find an alternative

   In the example calculation method, there are 3 alternative data that will be ranked.

   TABLE 13. ALTERNATIVE DATA

No	Employee name
1	Naufal
2	Dwi
3	Wawan

2) Then calculate the normalization of weights, calculate vectors and rank. The previous preference weights are:

   W = (2,3,4,4,5)

   Wj is the W index to j. So for W1 is 2, W2 is 3 and so on.

   Σwj is the sum of W is 2+3+4+4+5. So to fix the W1 weights it becomes:

$$W1 = \frac{2}{2+3+4+4+5} = -0.11111$$

$$W2 = \frac{3}{2+3+4+4+5} = -0.16667$$

$$W3 = \frac{4}{2+3+4+4+5} = -0.22222$$

$$W4 = \frac{4}{2+3+4+4+5} = -0.22222$$

$$W5 = \frac{5}{2+3+4+4+5} = -0.27778$$

Using the equations described previously, the vector values for each criterion are obtained as follows:

$S1 = (4^{-0.11111})\,(6^{-0.16667})\,(8^{-0.22222})\,(8^{-0.22222})\,(10^{-0.27778}) = 1.133126$

$S2 = (4^{-0.11111})\,(6^{-0.16667})\,(8^{-0.22222})\,(8^{-0.22222})\,(15^{-0.27778}) = 1.118945$

$S3 = (6^{-0.11111})\,(9^{-0.16667})\,(12^{-0.22222})\,(8^{-0.22222})\,(10^{-0.27778}) = 1.108697$

Calculating Preference (Vi) for ranking, the results are as follows:

$$V1 = \frac{1.133126}{1.133126 + 1.118954 + 1.108697} = 0.369007$$

$$V2 = \frac{1.118954}{1.133126 + 1.118954 + 1.108697} = 0.329701$$

$$V3 = \frac{1.108697}{1.133126 + 1.118954 + 1.108697} = 0.301293$$

From the calculation results above, the highest value is obtained by V1. In other words, V1 is an alternative that can be chosen as the best employee according to the criteria and weights set by the decision maker.

System Implementation

The main system page can be seen in Figure 4. On this page there are several main menus, namely homepage, company profile and login.

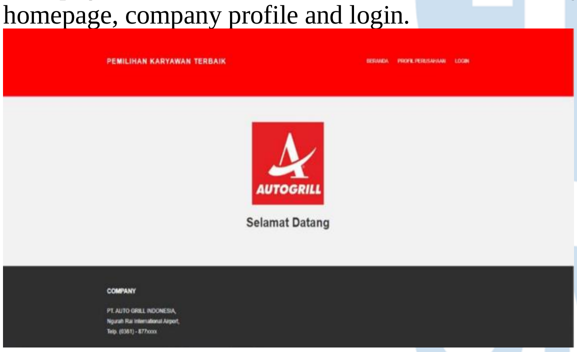


Figure 4. Main Page

Login Page

User access rights to enter the system are divided into four levels, namely: admin, outlet manager, operations manager, and employees.

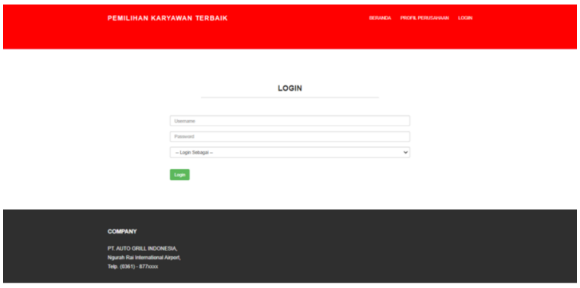


Figure 5. Login page

Admin Dashboard Page

The admin dashboard page is the main display of admin user access rights when successfully logged into the system. There is information on the number of users who have registered in the system.

Figure 6. Admin Dashboard Page

User Data Page

On this page, the admin can view user data information. Admin can also make changes to user data if an error occurs, and search user data.

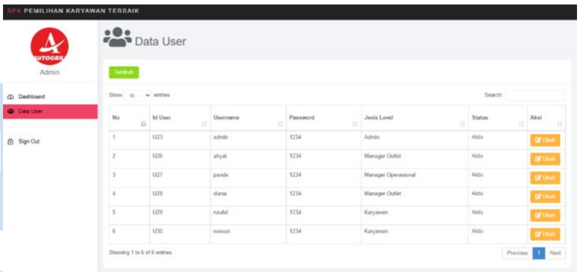


Figure 7. User Data Page

Add User Data Page

Admin can add new user data, by filling out a form consisting of name, username, password, type, outlet, and status. The condition for adding user data is that all fields cannot be empty. If it is empty, the system will give a warning that there is data that has not been filled in.

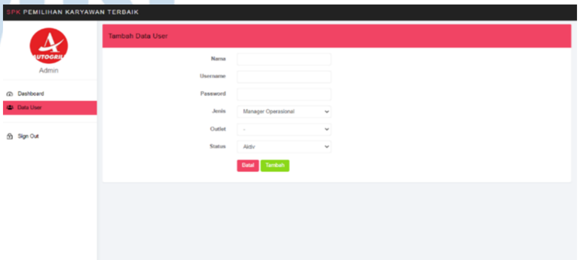


Figure 8. Page Add User Data

Outlet Manager Dashboard page

This page is the main display of the outlet manager when successfully logged into the system. Managers can view active employee data information that has been registered in the system. On this page there is a dashboard menu, employee data, criteria data, period, alternative data, score data, calculation process and calculation results.

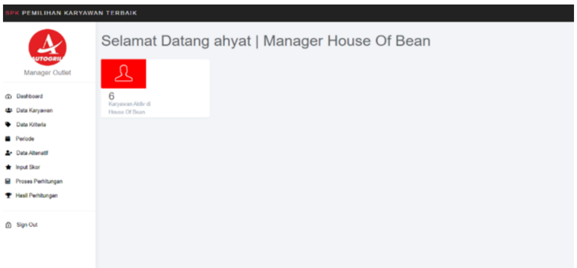


Figure 9. Dashboard Manager Outlet Halaman page

Employee Data Page

On the employee data page, there is a display of employee data information that has been registered in the system. On this page there are also add, change and search textbox buttons that can be accessed by the outlet manager.

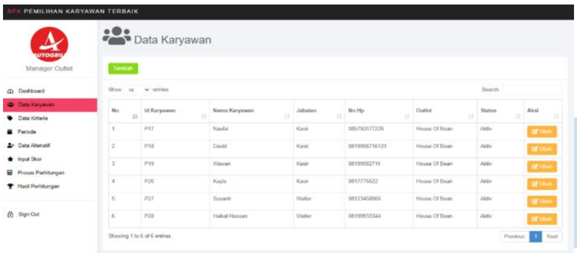


Figure 10. Employee Data Page

Add Employee Data Page

On the add employee data page, there is a form that must be filled out completely and correctly by the outlet manager to add new employee data to be stored in the system by clicking the add button.

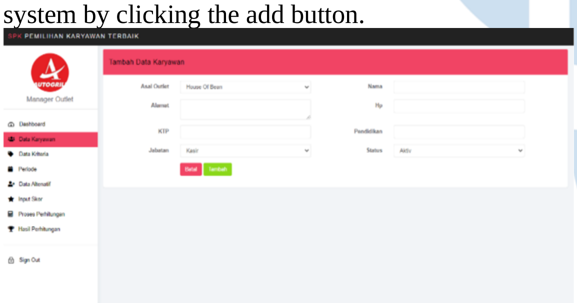


Figure 11. Add Employee Data Page

Change Employee Data Page

On the change employee data page, there is a form that must be completed with complete and correct data according to the new data that will be changed by the outlet manager then the manager clicks the edit button to save the new data into the system.

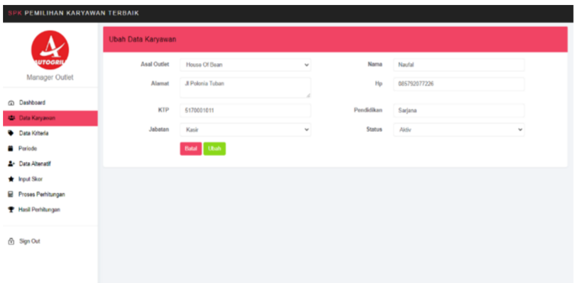


Figure 12. Change Employee Data Page

Criteria Data Page

On this criteria data page, there is information on criteria data that has been set by the company. On this page there are buttons add, change, delete, and a search textbox.

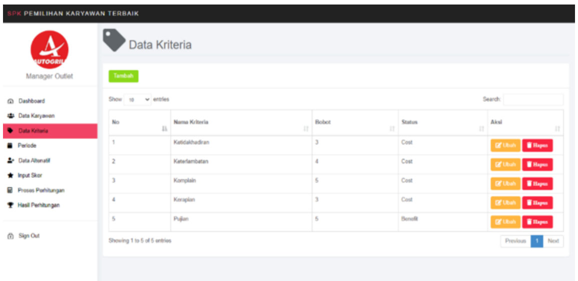


Figure 13. Criteria Data Page

Add Criteria Data Page

On the add criteria data page, the outlet manager can fill out the form on this page completely and correctly to add new criteria data to the system. The level manager can select the add button to save the criteria data into the system.

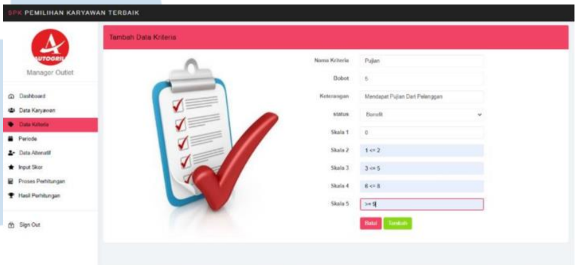


Figure 14. Add Criteria Data page

Period Data Page

On the data page of this period, there is information that displays the data for the period that is currently running on the system. On this page there is an add button to add period data and a save button to save period data.

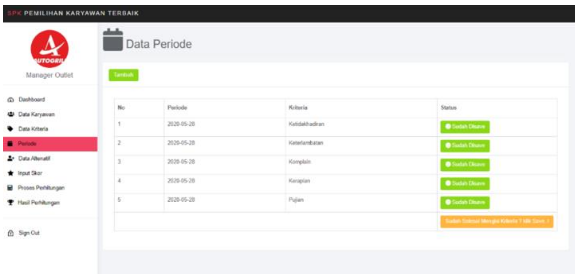


Figure 15. Period Data Page

Alternative Data Page

On the alternative data page, there is information that displays alternative data that has been saved into the system. On this page there are buttons add, delete, validate and textbox search.

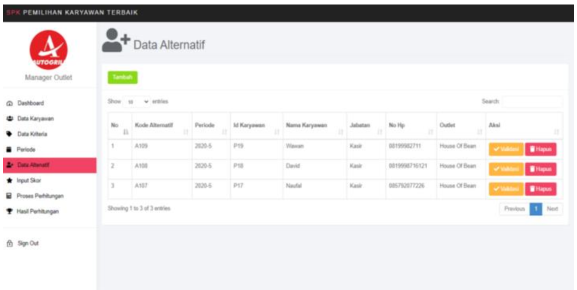


Figure 16. Alternative Data Page

Score Data Page

On the score data page, there is information that displays the score data that has been stored in the running system. On this page there is an add button, and a search textbox.

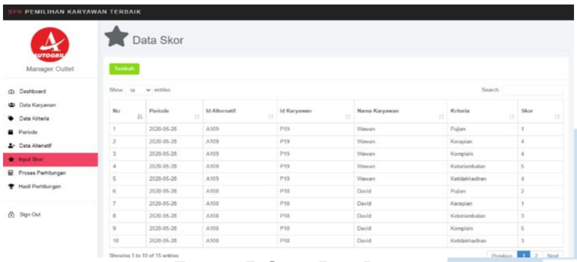


Figure 17. Score Data Pages

Calculation Process Page

On this page, before performing the calculation process the user is asked to select the calculation period according to the user's wishes. After that, you can proceed by selecting the calculation process button.

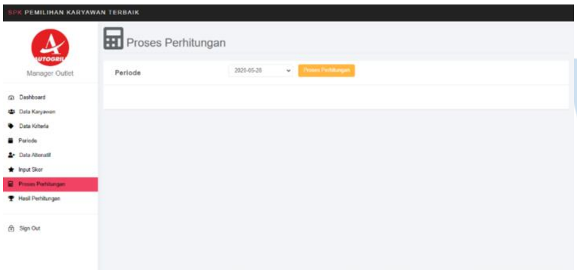


Figure 18. Calculation Process Page

Calculation Results Page

On the calculation results page, there is information that displays the results of the score calculation which has been completed by the outlet manager. On this page, you can also see information on the calculation results based on the month of the period selected by the outlet manager.

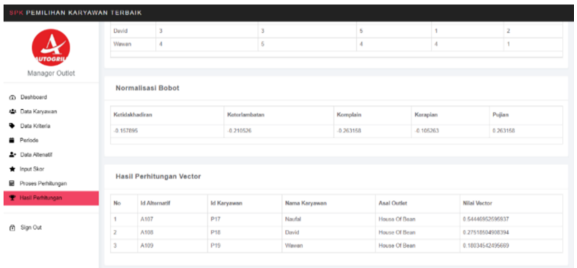


Figure 19. Calculation Results Page

Operations Manager Dashboard Page

This dashboard page displays data outlets that have been created by the previous operational manager. On this page there is a dashboard menu, data outlets, annual DSS and annual calculation results.

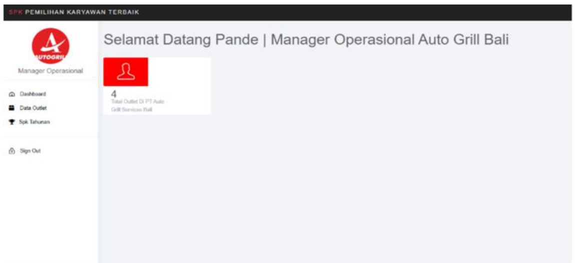


Figure 20. Operational Manager Dashboard page

Outlet Data Page

On the outlet data page, there is information that displays outlet data that has been previously stored by the operational manager. On this page there are add, change, and search textboxes.

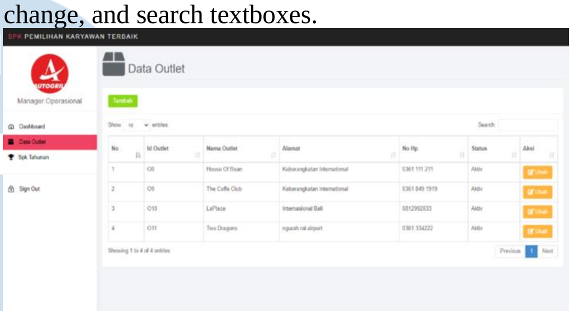


Figure 21. Data Outlet Page

Annual Participant Page

On the annual participant page, there is information on employee data for the best employee candidates from representatives of each outlet. On this page there is an assessment form that is filled out by the operational manager completely and correctly.

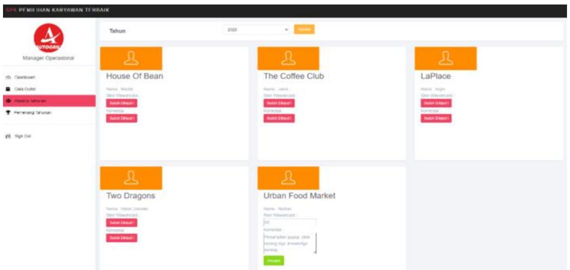


Figure 22. Annual Participant Page

Annual Winner Results Page

On this page, the operations manager can see the results of the annual winning scores based on the year selected by the operations manager.

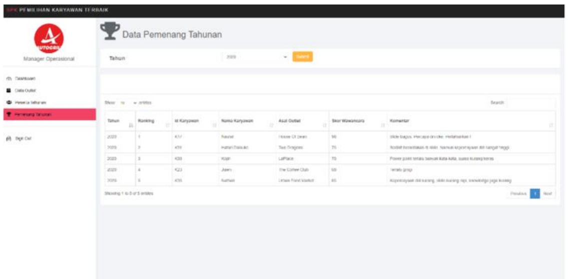


Figure 23. Annual Winner Results Page

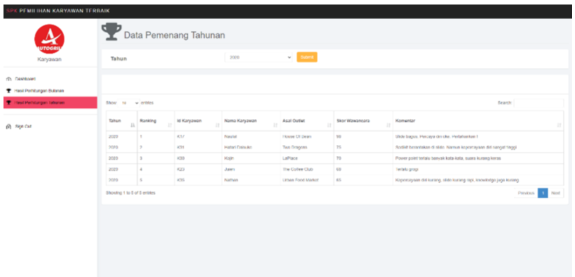


Figure 26. Annual Calculation Results Page

Employee Dashboard Page

On the employee dashboard page there is a dashboard menu, monthly calculation results, and annual calculation results that can be accessed by employees to see the results of the assessment from the outlet manager and operational manager based on the period to be selected.

Figure 24. Employee Dashboard Page

Monthly Calculation Results Page

On the monthly calculation results page, employees can see the assessment results from the outlet manager based on the selected period, but employees cannot change the results of calculations that have been made by the outlet manager, to prevent data manipulation.

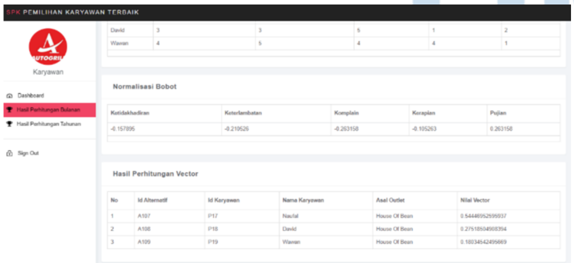


Figure 25. Monthly Calculation Results Page

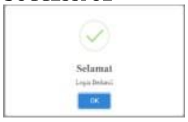
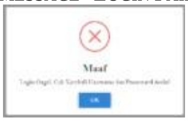
Annual Calculation Results Page

On the annual calculation results page, employees can see the results of the assessment from the operational manager based on the selected year, but employees cannot change the results of calculations that have been made by the operational manager, to prevent data manipulation.

Testing

The test method used is black box. In this test, all system functionality will be tested based on a pre-designed scenario.

TABLE 14. SYSTEM TEST RESULT

SCENARIO	TEST RESULT	*INFORMATION*
Enter Username and Password correctly	The system accepts Login access and displays the message "Login Successful" 	*Valid*
Wrongly fill in Username and Password	The system refuses Login access and displays the message "Login Failed" 	*Valid*
Empty one of the text-fields on the add user form	The system displays a message that the field cannot be empty 	*Valid*
Fill in all text-fields on the form add position data	The system displays a message that the data was successfully added berhasil 	*Valid*
Fill in the form to change user data correctly	The system displays a message that the data was successfully changed	*Valid*

SCENARIO	TEST RESULT	INFORMATION
		
SEARCH USER DATA BY KEYWORD	THE SYSTEM DISPLAYS DATA BASED ON THE SEARCHED KEYWORD 	*VALID*
EMPTY ONE OF THE TEXT-FIELDS ON THE FORM ADD JOB DATA	THE SYSTEM DISPLAYS A MESSAGE THAT THE FIELD CANNOT BE EMPTY 	*VALID*
FILL IN ALL THE TEXT-FIELDS ON THE ADD DATA OUTLET FORM	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY ADDED BERHASIL 	*VALID*
FILL OUT THE CORRECT OUTLET DATA CHANGE FORM	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY CHANGED 	*VALID*
SEARCH OUTLET DATA BY KEYWORD	THE SYSTEM DISPLAYS DATA BASED ON THE KEYWORDS YOU ARE LOOKING FOR 	*VALID*
EMPTY ONE OF THE TEXT-FIELDS ON THE ADD EMPLOYEE DATA FORM	THE SYSTEM DISPLAYS A MESSAGE THAT THE FIELD CANNOT BE EMPTY 	*VALID*
FILL IN ALL TEXT-FIELDS ON THE FORM ADD EMPLOYEE DATA FORM	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY ADDED BERHASIL 	*VALID*

SCENARIO	TEST RESULT	INFORMATION
FILL IN THE FORM TO CHANGE EMPLOYEE DATA CORRECTLY	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY CHANGED 	*VALID*
SEARCH EMPLOYEE DATA BY KEYWORD	THE SYSTEM DISPLAYS DATA BASED ON THE SEARCHED KEYWORD 	*VALID*
EMPTY ONE OF THE TEXT-FIELDS ON THE FORM ADD CRITERIA DATA	THE SYSTEM DISPLAYS A MESSAGE THAT THE FIELD CANNOT BE EMPTY 	*VALID*
FILL IN ALL TEXT-FIELDS ON THE FORM ADD CRITERIA DATA	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY ADDED 	*VALID*
FILL IN THE FORM TO CHANGE THE CRITERIA DATA CORRECTLY	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY CHANGED 	*VALID*
SEARCH CRITERIA DATA BASED ON KEYWORDS	THE SYSTEM DISPLAYS DATA BASED ON THE SEARCHED KEYWORD 	*VALID*

SCENARIO	TEST RESULT	*INFORMATION*
CONFIRM VALIDATION ADD ALTERNATIVE DATA	THE SYSTEM DISPLAYS A MESSAGE THAT THE VALIDATED DATA CANNOT BE CHANGED ANYMORE 	*VALID*
FILL IN THE FORM ADD ALTERNATIVE DATA CORRECTLY	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS SUCCESSFULLY ADDED 	*VALID*
PRESSING THE ALTERNATE DATA CLEAR BUTTON CORRECTLY	THE SYSTEM DISPLAYS A MESSAGE THAT THE DATA WAS DELETED SUCCESSFULLY 	*VALID*
FIND ALTERNATIVE DATA BASED ON KEYWORDS	THE SYSTEM DISPLAYS DATA BASED ON THE SEARCHED KEYWORD 	*VALID*
FILL IN ALL SCORE DATA BASED ON CRITERIA AND EMPLOYEE ID	IF ONE EMPLOYEE HAS FINISHED INPUTTING HIS SCORE, THE SYSTEM WILL CONTINUE TO INPUT THE NEXT EMPLOYEE'S SCORE AUTOMATICALLY 	*VALID*
SEARCH SCORE DATA BY KEYWORD	THE SYSTEM DISPLAYS DATA BASED ON THE SEARCHED KEYWORD 	*VALID*

Based on the test results, all system functionality can run well and in accordance with the design results.

IV. CONCLUSION

Based on the discussion and analysis of the results of system testing, it can be concluded that, the Implementation of the Weight Product (WP) Method on the Best Employee Selection at PT. Autogrill Services Indonesia has been successfully applied through a series of stages such as data collection, analysis, design, implementation and evaluation in building the system. This system functionality has been running well and the output results are in accordance with the design results.

The suggestions for development in further research include: adding backup and import database features on a regular basis as well as adding features for managers so that managers can input daily log data into the system.

REFERENCES

[1] W. Bangun, *Manajemen Sumber Daya Manusia*. Jakarta: Penerbit Erlangga, 2012.

[2] P. P. A. P. Indah, P. I. Rahmawati, and N. D. Andiani, “Pengaruh Sistem Penghargaan (Reward) Terhadap Kinerja Trainee Di Hotel Holiday Inn Resort Baruna Bali,” *J. Manaj. Perhotelan dan Pariwisata*, vol. 2, no. 1, pp. 41–54, 2019.

[3] A. A. A. P. Mangkunegara, *Evaluasi Kinerja SDM*. Tiga Serangkai, 2005.

[4] F. A. Prasasti, M. P. Hutagaol, and M. J. Affandi, “Pengaruh Penilaian Kinerja Terhadap Kinerja Karyawan Bagian Pemasaran Melalui Faktor Pendorong Motivasi (Studi Kasus: Bank Xyz Kcu Bekasi),” *J. Apl. Bisnis dan Manaj.*, vol. 2, no. 3, pp. 248–257, 2016.

[5] E. Turban, J. E. Aronson, and T.-P. Liang, *Decision support systems and intelligent systems (sistem pendukung keputusan dan sistem cerdas)*. Andi Offset, 2005.

[6] M. Rusli, S. Arifin, and A. Trisnadoli, “Pengembangan Sistem Pendukung Keputusan untuk Penentuan Lokasi Promosi Penerimaan Mahasiswa Baru,” *J. Komput. Terap.*, vol. 3, no. 1, pp. 11–18, 2017.

[7] S. Effiyaldi, “Analisis dan Perancangan Sistem Pendukung Keputusan Pemilihan Kendaraan Roda Dua Menggunakan Metode Simple Additive Weighting (SAW) (Studi Kasus : PT. Sinar Sentosa),” *J. Manaj. Sist. Inf.*, vol. 3, no. 1, pp. 911–921, 2018.

[8] L. Kristiyanti, A. Sugiharto, and H. A. W, “Sistem Pendukung Keputusan Pemilihan Pengajar Les Privat untuk Siswa Lembaga Bimbingan Belajar dengan Metode AHP (Studi Kasus LBB System Cerdas),” *J. Masy. Inform.*, vol. 4, no. 7, pp. 39–47, 2013.

[9] K. Eliyen and F. S. Efendi, “Implementasi Metode Weighted Product Untuk Penentuan Mustahiq Zakat,” *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 4, no. 1, pp. 146–150, 2019.

[10] E. Ningsih, Dedih, and Supriyadi, “Sistem Pendukung Keputusan Menentukan Peluang Usaha Makanan yang Tepat Menggunakan Weighted Product (WP) Berbasis Web,” *ILKOM*, vol. 9, no. 3, pp. 244–254, 2017.

[11] S. Susliansyah, R. R. Aria, and S. Susilowati, “Sistem Pemilihan Laptop Terbaik Dengan Menggunakan Metode Weighted Product (Wp),” *J. Techno Nusa Mandiri*, vol. 16, no. 1, pp. 15–20, 2019.

[12] Basri, “Metode Weighted Product dalam Sistem Pendukung Keputusan Penerimaan Beasiswa Prestasi,” *J. INSYPRO*, vol. 2, no. 1, pp. 1–6, 2017.

[13] A. Setyawan, F. Y. Arini, and I. Akhlis, “Comparative Analysis of Simple Additive Weighting Method and Weighted Product Method to New Employee Recruitment Decision Support System (DSS) at PT. Warta Media Nusantara,” *Sci. J. Informatics*, vol. 4, no. 1, pp. 34–42, 2017.

[14] A. S. Purnomo and A. F. Rozi, “Comparative Analysis Uses Weighted Product (WP) and Simple Additive Weighting (SAW) Methods in the Best Graduation Selection System,” *Sebatik*, vol. 23, no. 1, pp. 44–52, 2019.

[15] E. W. Fridayanthie, N. Khoirurrizky, and T. Santoso, “Sistem Pendukung Keputusan Pemilihan Karyawan Terbaik Menggunakan Metode Weighted Product,” *Paradig. - J.*

Komput. dan Inform., vol. 22, no. 1, pp. 41–46, 2020.

[16] M. Adnan Farizhi and A. Diana, “Sistem Pendukung Keputusan Pemilihan Karyawan Terbaik Untuk Promosi Jabatan Dengan Metode WP (Weight Product),” *Pros. Semin. Nas. Teknoka*, vol. 5, no. 2502, pp. 9–18, 2020.

[17] A. A. Arwaz, T. Kusumawijaya, R. Putra, K. Putra, and A. Saifudin, “Pengujian Black Box pada Aplikasi Sistem Seleksi Pemenang Tender Menggunakan Teknik Equivalence Partitions,” *J. Teknol. Sist. Inf. dan Apl.*, vol. 2, no. 4, p. 130, 2019.

[18] D. Febiharsa, I. M. Sudana, and N. Hudallah, “Uji Fungsionalitas (BlackBox Testing) Sistem Informasi Lembaga Sertifikasi Profesi (SILSP) Batik Dengan AppPerfect Web Test Dan Uji Pengguna,” *JOINED J.*, vol. 1, no. 2, pp. 117–126, 2018.

FastText Word Embedding and Random Forest Classifier for User Feedback Sentiment Classification in Bahasa Indonesia

Yehezkiel Gunawan[1], Julio Christian Young[2], Andre Rusli[3]
Department of Informatics, Multimedia Nusantara University, Tangerang, Indonesia
[1]yehezkiel.gunawan@student.umn.ac.id, [2]julio.christian@umn.ac.id, [3]andre.rusli@lecturer.umn.ac.id

Accepted 14 June 2021
Approved 18 June 2021

***Abstract*— User feedback nowadays become a platform for software developer to identify and understand user requirements, preferences, and user's complaints. It is important for the developer to identify the problem that exist in user feedback. According to software growth, user amount also growth. Read and classify one by one manually are wasting time and energy. As the solution for the problem, sentiment analysis system using Random Forest Classifier which use word embedding as the feature extraction is made to help to classify which feedback is positive, neutral, or negative. Random Forest Algorithm is chosen because it gives the best performance, even its need the larger resources. Furthermore, with word embedding, the words which has semantic or syntactic similarities will be detected. Word embedding does not need stemming and stop word removal, so the context of the sentences keep remains. This research is made to implement word embedding to classify sentiment of user feedbacks using Random Forest Classifier. 70.27% accuracy, 80% precision, 54 recall and 54% F1 score is reached when BYU dataset (200 dimension) as embedding dataset with the train and test ratio 80:20.**

Index Terms— ***Bahasa Indonesia, Random Forest, Word Embedding, NLP, user feedback***

I. Introduction

In software engineering, there is a term called 'requirements engineering'. It means a process which the requirements for a software are assembled, documented, and managed as long as the software engineering process [1]. Interpreting and understanding the purpose, stakeholder's requirements and trust are the main focus of requirement engineering [1].

It was shown in [2] that nowadays users or clients are involved in the software engineering process, so the developers will know the needs, preferences, and problems which users experienced. Software developers must understand the issues or problem that appear from the app which they developed, like bugs, unwanted feature, and adding the new feature which accurate and on time in the future [3]. Users who experience some issues when they use an app can send some feedbacks that can reflect their experience when they use it, then the feedbacks can be considered by the developers to improve the app quality [3].

When the scope of the application are become greater, it is much challenging to identify user feedbacks issues [4]. Modern problems require modern solutions, so a recommendation system or sentiment classification needs to be made regarding to the growth of user feedbacks [4]. There are more than a hundred or even thousand user feedbacks have been sent in one day. Checking users' comment one by one can be exhausting and wasting time. That is the reason why automated sentiment analysis is needed to analyze and generalize the user's feedback with the sentiment analysis technique.

A method called CRISTAL is introduced by Palomba *et al* [4] which can detect the impact caused by the informative user comment on the changes or updates of the app's source code. The research is done to determine how impactful the app review on the software development process [4]. Its result stated that 49% of developers will consider the informative user comments or feedback for their next updates. It also considered very impactfully on the app success because the increase of the rating and positive feedback from the users directly compared with the fulfillment of the requirements based on the user review. This result can strengthen the reason why we need a system to recommend or classify the user feedbacks to help the developer on developing their app [4].

Sentiment analysis is the process that learn people's opinion, sentiment, emotion, rating, and gesture on an entity [5]. There are a lot of activities that are related to the sentiment analysis process and even harder to separate it because there are a lot of aspects, and one of them is sentiment classification. Sentiment classification is based on the idea that text can be the expression of a person's opinion on an entity and trying to predict what kind of sentiment that can be resulted [6]. Machine learning in general can be used to classify the sentiment and give the good accuracy.

As in [7], sentiment analysis is an automated process to mine and classify opinion, view, emotion, and sentiment from the text dataset which are not structured for machine language and computer programming. In sentiment classification, text can be classified to several labels, for example positive or negative.

Nowadays, people express their opinion with their language which tends to be ambiguous and complicated words [8]. Commonly, there are related words one each other and it often seems similar. To help to solve the problem, there is a method called word embedding. It is a kind of word representation that make the words which have similarity can be understood by the machine learning algorithm [9]. Technically, the input words will be mapped into number vector using neural network, probability model, or the dimension reduction on the word co-occurrence matrix.

It is stated in [10] that word embedding is considered can learn the word vector with high quality from a big dataset. Instead of that, the existing vocabulary that has been made from the pre-trained model of the word embedding also considered detecting the word similarity both semantically or syntactically. It can help the machine to recognize the similarity which exists in the dataset.

Currently, two big platforms that provide the apps choice and review, are Google Playstore and Apple Appstore. The user feedback dataset from these platforms is considered because when a user wants to give some feedback to an app, he or she must have an account so it will allow the user to give some review. For example, when a person wants to give some feedback to an app in Google Playstore or an Apple Appstore, so that person must have a Google or Apple account to allow him or her to write the feedback on their smartphone. Besides that, some developers can set the app which they made to allow the user to give some review within a certain period, so that can be confirmed if the user who gives the feedback is the user of the app.

This research will use *word embedding* as the feature extraction and Random Forest Classifier to classify the user feedback's sentiment. This research aims to study, analyze, and implement the Random Forest Classifier to analyzing the sentiment of application user feedbacks in Bahasa Indonesia. The paper is organized as follows. In the following section, we review some related works of ours. Then we present a brief overview of some research in sentiment analysis.

II. Related Work

A comparative study has been done by [8] on few machine learning algorithms used for classification. Those are Naïve Bayes, Max Extropy, Boosted Trees, and Random Forest. The result of the research is Random Forest has the best performance with the high simplicity even it requires more resources.

Following similar trends, several works of literature can also be found working on sentiment analysis of documents. However, most of them focus on analyzing tweets from Twitter. Like similar observation done by Vora, Khara, and Kelkar [9], they used different word embedding methods as the feature extraction to classify the sentiment of English tweet. For the classification algorithm, they used Random Forest Classifier. The result shows that when they used FastText with 300 dimensions as the feature extraction, the accuracy reached 91%.

Based on those previous works on sentiment analysis, natural language processing, and the requirements engineering activity in the software engineering field, this research focuses on adapting [9] research but with different dataset. Our research aims to implement FastText and Random Forest Classifier to classify the sentiment of application user feedback in Bahasa Indonesia and measure the performance.

III. Research Methods

A. *Requirements Engineering*

Requirement engineering is the process of gathering, analyze, documenting, and managing the requirements needed for software development [1]. It always related to determine and understand the purpose, requirements, and even what the stakeholder trust [1].

In requirement engineering, some things need to be done, those are feasibility study, finding the requirements (gathering information and analyze), convert the requirements into the standard or specification, and ensure that the requirements are based on the user's need (validation) [11]. In reality, this process is iterative and interleaved [11].

According to [11], in the requirement gathering process, the client is involved to determine the scope of the app, what services will exist in the system and the operational limits of the system. It may be involved the user, manager, engineer, and the people who will maintain the system. In all system, the requirements can be changed, the people who involved, developed a better understanding of what are they want in the app when it's released, like the changes on the hardware, software, or the system environment [11]. The understanding and control process of changes on the system requirement is called requirement management [11]. Besides that, the feedback from the user can cause changes to the requirements [2].

The software development activity [11] is not stopped when the software has been released. But it continues throughout its lifetime. There are five steps of the software lifetime, those are the initial development, the software engineer build the first version of the app or system, then evolve, the ability and functionality of the system are expanded to fulfill the user preferences, then servicing, the system is fixed from the bug, issues, and update the functionality, then phaseout, the system owner decided to stop the servicing process and make income from the system for a long time, and the last step is closedown, the owner takes down the system from the market and direct the users to the new system.

B. Natural Language Processing

Natural language processing (NLP) is a computer science field dealing with human language processing in either text or speech [12]. In this research, preprocessing of user feedback includes the punctuation, remove special character, lowering case, and tokenization. All that preprocessing method is done by the help of library string and re.

C. Word Embedding

It is shown in [13] that the computer can learn the character input with feature extraction. Every kind of feature is taken from the dataset, then the machine will learn it. In this research, the feature which will be extracted is the word similarity.

There is a problem that has to be faced. The computer only can read the numbers. If the received objects are words, so it needs to be converted into the numeric vector which represents each word. Because of that, word embedding can be used as a feature extraction to learn the similarity between words. It uses a neural network model to learn the words [13].

Word embedding represent the words into vectors. For example, there is the sentence Word Embedding are Word Converted into numbers. First, it will make a dictionary to contain it, and the dictionary is [Word, Embeddings, are, Converted, into, number]. With one-hot encoding, it will represent the vector where 1 represent the position of a word. The vector representation of word numbers from that example is [0,0,0,0,1].

This method learns the vector representation of the constant vocabulary which exists in the corpus or dataset. It also uses neural network model for the task like document classification or with unsupervised learning using document statistics [13]. In general, three common models are often used to do word embedding, those are Latent Semantic Analysis (LSA), Word2Vec, and GloVe.

D. FastText

This research will use FastText as the word embedding method. FastText is the library from Facebook to do the word embedding method [14]. FastText is a newest version from Word2Vec which Google made. Actually, both of them can be used to determine the word similarity.

According to the Indonesian Dictionary, semantic means the language structure related to the meaning of an expression or the structure of the meaning of a talk or text. Meanwhile, the syntactic comes from the 'syntax' word adopted in English which is the structure or writing. In other words, if there is a word similarity syntactically, it means there is a similarity in the writing structure.

The input words will be represented into the vector and placed in such a way so the words which have the similar meaning will appear close by, meanwhile, the opposite will appear far from the vector. The main difference between FastText and Word2Vec is FastText can process the input words which not exist in the vocabulary or out of vocabulary words.

Like Word2Vec, there is two architecture on FastText, those are Continuous Bag of Words (CBOW) and Skip-Gram. CBOW predicts the current word (as a target) from the context (as an input) around it. Meanwhile, the Skip-Gram uses the current word (as an input) to predict the context (as a target). The visualization of CBOW and Skip-Gram can be seen in Fig.1

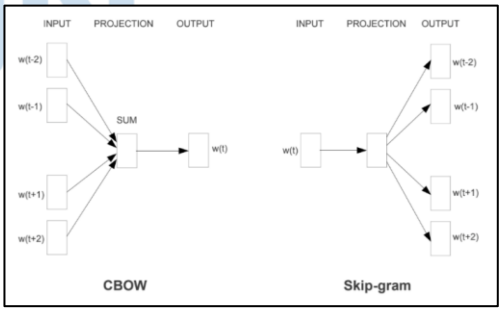


Fig. 1. The concept of CBOW and Skip-Gram [10]

In the Fig.2, there is an example which given the input words "the best revenge is massive success" and there is a forward-backward training with the CBOW architecture. Assume that w(t-2) = "the", w(t-1) = "best", w(t+1) = "is", w(t+2) = "massive" as input and w(t) = "revenge" as target.

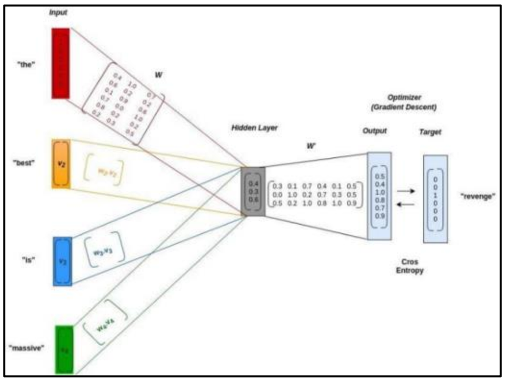


Fig. 2. Ilustration of forward-backward training CBOW [10]

[14] said it requires a special dataset to get the expected result. [14] makes a term for the dataset which used for the training process using FastText as an embedding dataset. The embedding dataset trained using FastText will produce a vocabulary that consists of the vocabs that can be used to detect the word similarity. The result of the training of the process can be called a pre-trained model.

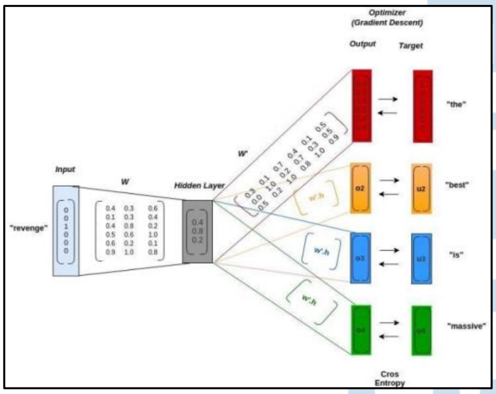


Fig. 3. Ilustration of forward-backward training Skip-Gram[10]

E. Decision Tree

Decision Tree is a machine learning method that learns and take decision with the functional target that has discrete values [15]. This technique can be represented with a group of if-then rules so it will be understandable. Each tree consist of the leaf and branch. Each leaf reflect the group attribute which will be classified and each branch represent the values which taken by the leaf.

There are three parts of decision tree like the Fig.4 [16]. First is the root node which is the first node and there is no input branch in this node. The second is internal node which has branch, and just has one input and minimum two output. The last is leaf node that is the node which has just one input and no output.

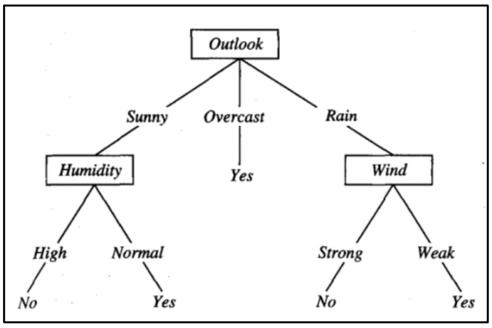


Fig. 4. Decision Tree Concept [6]

The attribute selection [16] can be done by the Gini Index process. Gini index is a metric that measure how often the misclassification happened [15]. In this process, the smallest Gini Index value will be selected and be the root node or internal node. As it said in [15], (1), (2), and (3) can be used to determine Gini Index.

$$Gini\ (D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (1)$$

$$Gini_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{D} Gini(D_i) \quad (2)$$

$$Gini\ Gain\ (A) = Gini(D) - Gain_A(D) \quad (3)$$

According to [16], there are five steps to make a decision tree using Gini Index. First is determine the class or label which will be the root in the tree using (1). All the lable's probability with the constraint that has determined will be calculated for the probability and squared. The second is calculate the Gini Index on every attributes or features in the dataset using (2). All the label's probability, in general, will be multiplied by each column of the Gini Index and calculated.

The third is to choose the lowest Gini Index. The feature with the lowest Gini Index will be the root of an internal node in the decision tree. Then, in each branch, do the recursive way from the first step until the leaf or the Gini Index on the branch is zero. Then the last thing is to calculate the Gini Gain with (3) to determine the difference from (1) using Gini Index from each first branch.

F. Random Forest Classifier

Random Forest is an algorithm of machine learning that has ability to do the regression or classification task [17]. This algorithm consists of many decision trees which randomly selected from the subset of the training set. The classification of Random Forest is the accumulation of the votes which decision trees did. The process can be seen in Fig.5.

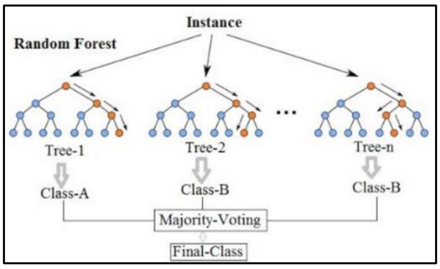


Fig. 5. Random Forect Classifier Structure [9]

According to [18], bagging or bootstrap aggregation is the common technique used for Random Forest Algorithm. The bagging technique is used to reduce the variance of the prediction function which has been estimated. This method is considered effective for the data which has big variance and procedure which has a low bias like trees. In the case of classification, each tree will produce the prediction result which will calculate the majority result or in other words majority vote [18].

The tree-making process [19] can be done by Gini or Entropy. Gini is preferred for the attribute which has continuity, meanwhile, entropy is commonly used for a discrete attribute that exists on each label. The tree-making process can be seen in (4), (5), and (6) which is equal to the decision tree formula [20]. A decision tree can be made recursively based on the tree amount which has been determined. [19] also said that commonly the optimal tree amount made for the classification process is $\sqrt{p}$ and for regression is $\frac{p}{3}$ which p is the predictor amount that will be used. The output of each tree for the classification process will be submitted and there will be a majority vote process like (4) [18].

$$\hat{C}_{rf}^{B}\ (x) = majority\ vote\{\hat{C}_b(x)\}_1^B \qquad (4)$$

IV. Result and Discussion

A. Dataset

In work conducted in this paper, we use user feedbacks from another research [9]. There is a total of 553 feedbacks, which is contains of 259 positive, 241 negative, and 53 neutral feedbacks.

B. Performance Evaluation

There are several scenarios are conducted to determine the best configuration for classifying the feedbacks' sentiment. An experiment with two training-testing ratios, 70:30 and 80:20, is conducted. First we try to use the feedback as it is (Scenario 1). Furthermore, we try to upsample the neutral feedbacks (Scenario 2) and use the GridSearchCV library to find the best hyperparameter (Scenario 3).

In each scenario, we used different embedding dataset to make the FastText pre-trained model, such as user feedbacks of BYU and Tokopedia app from Playstore (200 & 300 dimension), and SentimentAppReview dataset itself. We also used the original FastText pre-trained model from its official website as the feature extraction.

In the experiment following Scenario 1, the best performance is reached when the embedding dataset is BYU (200 and 300 dimension) and the training-testing ratio is 80:20. The accuracy is 70,27%, the precision is 80%, the recall reaches 54%, and the F1 score is same as the recall. This scenario can result the precision, recall, and F1 score equally. In all scenarios, the performance to detect neutral feedbacks is lower than the others because it is unbalanced. The neutral feedbacks from the dataset is just 53 of 553 row data.

In general, if there is some addition on the test set and subtraction on a train set, the accuracy possibly increases. But, in this research, the 80:20 ratio is better than 70:30. Not only on accuracy, but also the precision, recall, and F1. Apart from the difference of train and test set ratio, the FastText pre-trained model which has chosen just give a small impact on the evaluation result.

Both 80:20 and 70:30, in general, seem hard to detect the neutral label. This is caused by the lack of neutral labels in the dataset. It just 53 if 553 rows. That is why the classifier model is hard to detect the neutral label due to the underfitting issue.

From the requirement engineering side, user feedback can be used to learn the needs, complaints, and user reviews. The negative feedbacks generally contain information that can be useful for future releases, like the bug report. The positive feedbacks usually contain the expression of thanks and user's satisfaction when they use the feature. So, it is important to detect the feedback which has negative or positive sentiment.

The precision calculates the prediction for the true positive to all positive label, that is why the precision score is the ideal measurement when the false positive has the high cost. In this case, false positive happens when feedback whose label is negative but predicted as positive. Meanwhile, the false-negative happens when feedback whose label is positive but it predicted as negative.

In Scenario 2, we try to up-sample the neutral feedbacks. The result is no one of the accuracy reached 70%. The best performance is reached when we used Tokopedia (200 dimension) as embedding dataset and the training-testing ratio is 70:30. The accuracy is 69,63%, the precision is 76%, the recall is 63%, and the F1 score is 64%. According to the result of this scenario, up-sample can increase the performance from the precision, recall, and F1 score side, but it decreases the accuracy.

Scenario 2 can produce the best result when using Sentiment App Review (200 dimensions) as the FastText pre-trained model and with the ratio of 80:20. In Table 1, we can see that the result of the training set is nearly perfect. There is still misclassification on the neutral label due to the lack of the neutral label data on the dataset and the up-sampling process which is done is not too big. Meanwhile, the data which has positive and negative label relatively give the precise prediction result because the amount of both is quite bigger and balanced.

TABLE II. THE BEST MODEL PERFORMANCE OF SCENARIO 2

Real Label \ Predicted	Negative	Neutral	Positive
Negative	192	0	1
Neutral	0	48	0
Positive	0	0	207

Based on the result and analysis of Scenario 2, upsampling is proven to increase the performance of the precision, recall, and F1 score, but it decreases the accuracy score even it is not too significant.

In Scenario 3, we try to use GridSearchCV to find the best hyperparameter for the Random Forest Classifier. GridSearchCV is usually used to tune the hyperparameter. The method that used in this case is Cross-Validation. It is the statistical method to evaluate the model or algorithm where the data is separated into two subsets, which are training and test. Like the train_test_split, but in Cross-Validation, the data is split into the fold which has been determined. In each Fold, there are train and test sets. Its process is iteratively done until all data from the dataset is included in the fold.

With the help of the GridSearchCV library, the Random Forest Classifier can tune its hyperparameter automatically based on the initial parameter that has been determined to find the best hyperparameter that can produce the best result. The hyperparameter tuned by the GridSearchCV in this case is the amount of leaf, tree, and max features. The tuning technique is done by initializing the input parameters in an array which will be tried on one by one by the library.

In our case, the best parameter for the Random Forest Classifier is 81 for the n_trees, 15 for the n_leafs, and log 2 for the max_features. The best result is reached when BYU (200 and 300 dimension) user feedbacks are used as the embedding dataset and the training-test ratio is 70:30. The accuracy is 71.68, the precision is 48%, the precision is 53%, and the F1 is 50%. This result if compared with Scenario 1, it is lower than Scenario 1. This scenario cannot detect the neutral feedbacks better than Scenario 1.

V. CONCLUSION AND FUTURE WORKS

In this research, we used FastText pre-trained model as the feature extraction and Random Forest Classifier to classify the sentiment of the user feedbacks. With this approach, we used the SentimentAppReview dataset from [12] which then is classified into positive, neutral, and negative label.

In conclusion, the implementation of word embedding to classify user feedback using Random Forest Classifier is successfully done. FastText pre-trained model which is used in this research is made of the user feedback dataset from the Tokopedia and ByU App at the Google Playstore.

Based on the results before, the best result is reached when we used BYU user feedbacks dataset as embedding dataset to make FastText pre-trained model as the feature extraction and the training-testing ratio is 80:20. The accuracy is 70,27%, the precision is 80%, the recall is 54%, and F1 score is 54%.

In the future work, we hope that there is a larger scope of the dataset so it maybe increases the performance with the same method. The more data we can get, hopefully, the better the result that we can get.

Furthermore, to make the more various results we can try different up-sampling or down-sampling methods. Hopefully, with this variance of methods, the result of the experiments can be more variance too.

Trying another word embedding library or feature extraction methods such as Word2Vec and GloVe is not a bad idea. With the help of the Gensim library, that thing can be made easier and hopefully give better performance.

Trying another cross-validation method or library also can be a good thing. For example, using the RandomizedSearchCV library to do the cross-validation method. With this, maybe the hyperparameter tuning can be more variative and give a better result.

If this classification method is combined with synonym extraction, we think it will increase the performance significantly. It will classify the word which has the same meaning with other word into one class more accurately. It will decrease the misclassification also.

REFERENCES

[1] A Van Lamsweerde (2009). Requirements engineering : from system goals to UML models to software

specifications. Chichester, England ; Hoboken, Nj: John Wiley, Cop.

[2] Morales-Ramirez, I. (2013). On Exploiting End-User Feedback in Requirements Engineering. [online] Available at: https://www.researchgate.net/publication/237064827_On_Exploiting_End-User_Feedback_in_Requirements_Engineering

[3] Gao, C., Zeng, J., Lyu, M.R. and King, I. (2018). Online app review analysis for identifying emerging issues. Proceedings of the 40th International Conference on Software Engineering - ICSE ’18.

[4] Palomba, F., Linares-Vasquez, M., Bavota, G., Oliveto, R., Di Penta, M., Poshyvanyk, D. and De Lucia, A. (2015). User reviews matter! Tracking crowdsourced reviews to support evolution of successful apps. 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME).

[5] Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4).

[6] Serrano-Guerrero, J., Olivas, J.A., Romero, F.P. and Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services.Information Sciences, 311, page.18–38.

[7] Fiarni, C., Maharani, H. dan Pratama, R., 2016, May. Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. In Information and Communication Technology (ICoICT), 2016 4th International Conference on (page. 1-6). IEEE.

[8] Gupte, A., Joshi, S., Gadgul, P. and Kadam, A. (2014). Comparative Study of Classification Algorithms used in Sentiment Analysis. International Journal of Computer Science and Information Technologies, 5(0975–9646), page.6261, 6264.

[9] Vora, P., Khara, M. and Kelkar, K. (2017). Classification of Tweets based on Emotions using Word Embedding and Random Forest Classifiers. International Journal of Computer Applications, 178(0975 – 8887).

[10] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. [online] Tersedia di:https://arxiv.org/pdf/1301.3781.pdf [Accessed 6 Feb. 2020].

[11] Sommerville, I. (2016). Software engineering. Harlow: Pearson Education.

[12] Pamungkas, E.W., Putri and Prasetyo, D.G. (2017). Word Sense Disambiguation for Lexicon-Based Sentiment Analysis. In: Proceedings of the 9th International Conference on Machine Learning and Computing. ACM, pp.442–446.

[13] Abdullah, A. (2018). Word Embedding. [online] Available at: https://rpubs.com/ahmadhusain/wordembedding [Accessed 8 Jan. 2020].

[14] Rehurek, R. (2019). Gensim: Topic Modelling For Humans. [online] Radimrehurek.com. Available at: https://radimrehurek.com/gensim/auto_examples/tutorials/run_fasttext.html#sphx-glr-auto-examples-tutorials-run-fasttext-py [Accessed 6 Feb. 2020].

[15] Mitchell, T.M. (2017). Machine learning. New York: Mcgraw Hill.

[16] Han, J., Kamber, M. and Pei, J. (2012). Data Mining Concepts and Techniques. 225 Wyman Street, Waltham, MA, 02451, USA: Morgan Kaufmann Publishers.

[17] Eletter, S., Yasmin, T., Elrefae, G., Aliter, H., & Elrefae, A. (2020). Building an Intelligent Telemonitoring System for Heart Failure: The Use of the Internet of Things, Big Data, and Machine Learning. 2020 21st International Arab Conference on Information Technology (ACIT). https://doi.org/10.1109/acit50332.2020.9300113

[18] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The elements of statistical learning, second edition: data mining, inference, and prediction. New York: Springer

[19] Sieling, G (2014). Decision Trees: “Gini” vs. “Entropy” criteria. Minnesota.

[20] Breiman, L. (2001). Random Forests. [online] 45. Tersedia di: stat.berkeley.edu/forests_V3.1.pdf [Accessed 8 Mar. 2020

ISSN 2085-4552

Spam Filtering on User Feedback via Text Classification using Multinomial Naïve Bayes and TF-IDF

Septiyan Mudhiya Sadid[1], Julio Christian Young[2], Andre Rusli[3]
Department of Informatics, Multimedia Nusantara University, Tangerang, Indonesia
[1]septiyan.mudhiya@student.umn.ac.id, [2]julio.christian@umn.ac.id, [3]andre.rusli@lecturer.umn.ac.id

Accepted 18 June 2021
Approved 12 September 2021

***Abstract*— User feedback could give a developer information on what should be fixed or should be improved. But there are many users feedback that is a spam. In user feedback, spam contents are more likely to be inappropriate feedback, feedback that is not feedback, just some random comment or even a question. Reading and choosing feedback manually could be costly, especially in terms of time and energy. Therefore, this research focuses on building a spam filtering model using Multinomial Naïve Bayes that implement a TF/IDF approach to detect spam automatically. For text classification, Multinomial Naïve Bayes proved on having better speed and having good performance. With TF/IDF, a word that highly occurred in many documents has less impact than others so it could help increasing performance from an imbalanced dataset. This research aims to implement Multinomial Naïve Bayes for spam filtering in user feedback and to measure the performance of the model. The best performance of this classifier was obtained when using the up-sampling method and typo corrector with a 70:30 ratio of train and test set resulting in 89.25% for accuracy, 45% for precision, 56% for recall, and 50% for F1-Score.**

Index Terms— Spam filtering, Multinomial Naïve Bayes, User Feedback, TF/IDF, Requirements Engineering

I. INTRODUCTION

In software development, user requirements is a useful component that developers needed. In order to correctly analyse and define the users' problem and needs, requirements engineering has become an important step in a software development lifecycle. Requirements engineering is a branch of software engineering which deals with problems that involve goals, functions, and restrictions on software systems [1]. Requirements engineering help software developer to have a better understanding about the problems. One of many ways to do requirements engineering is to directly involve the users in the development process. In the case in which the users of a software product is the public, it would be difficult to get all the users to be involved directly in the requirements specification and definition process. Many software developers rely on the users feedback gathered from many different sources online. For example, a mobile app developer in the Android environment would gather a lot of user feedback from the Google Play Store, and iOS environment from the Apple App Store. Another example is the web application of an e-commerce platform, the developers usually provide comment sections in order to gather feedbacks from their users, albeit content-related, product-related, or other categories of feedback.

User feedback is defined as all information obtained from users about whether they are satisfied or not with a product or service [2]. Users will give feedback if they are satisfied or there is a problem when they use the software. Even though user feedback has a lot of benefits for software development, some problems could occur such as spam in user feedback. Spam is an activity to send a message to other people with an electronic device continuously without the consent of the other party [3]. Spam is usually targeting random people. Spam messages always have the same characteristics, therefore it is not too hard to distinguish. This is particularly a serious problem for software products with thousands or even millions of user feedback. Manually reading each of the feedback could take time and effort which otherwise could be used for other development activities. It is more time-wasting if the feedback themselves are actually spam messages.

In user feedback, spam is more like inappropriate feedback, which is the feedback that is not actual feedback such as insulting feedback, fake review, etc. Inappropriate feedback is categorized as spam because that feedback is not useful for requirements engineering. Therefore, a classification system needs to be made to predict which feedback is spam. This research will use TF-IDF along with Multinomial Naive Bayes to make a classification system to detect spam in user feedback.

II. Related Work

Multinomial Naive Bayes is an improvement from Naive Bayes Classifier. Naive Bayes itself is a classification method that uses Bayes Theorem where each feature is assumed as independent to each other [4]. According to research from [5], the Multinomial Naïve Bayes algorithm runs 2 to 6 times faster than the Support Vector Machine. Multinomial Naive Bayes do better performance in 9 out of 13 models that [5] have. According to research from [6], the Multinomial Naïve Bayes algorithm has 89.58% accuracy in the letter classification system.

When dealing with text classification, it is necessary to do feature extraction before the classification. One of many popular feature extractions on text classification is TF-IDF and Doc2Vec. According to [7], TF-IDF has a better performance than Doc2Vec with 95% accuracy on Logistic Regression and 73.62% accuracy on Naïve Bayes. From the previous works, experiments related to the use of Multinomial Naïve Bayes and TF-IDF, which have been proven to be effective in other cases, for filtering spams in user feedback is still rare to be found. The main contributions of this research is to implement TF-IDF along with Multinomial Naïve Bayes to classify user feedback in Bahasa Indonesia and measure the performance of the system. Based on the results, we hope to be able to give more options to the developers to better their requirements engineering activities, especially when dealing with user feedback. Hopefully, by automating the spam filtering activity, more of the developers time and effort could be spent on other productive and creative activities to improve their products.

III. Research Methodologies

A. Dataset

Dataset used in this research is user feedback from tiket.com app in Bahasa. Dataset is gathered using an automatic scrapping extension on the google play store. Then, the dataset is labelled by 5 of our colleagues with a majority vote. The total dataset used in this research is 900 data with 810 data of Ham and 90 data of Spam. The distribution of the dataset can be seen in Fig. 1.

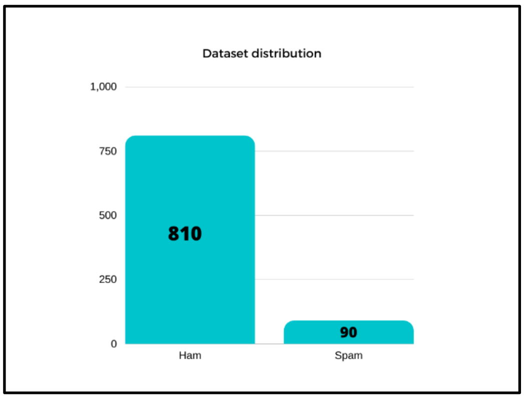


Fig. 1. Dataset distribution

B. Preprocessing

Preprocessing involved in this research is consist of Tokenizing, Filtering, and Stemming.

Tokenizing separates each sentence into an array of words. For example, the sentence “There are 2 winners of this game” will be tokenized like (‘There’, ’are’, ’2’, ’winners’, ’of’, ’this’, ’game’).

Filtering is a process that removes a bunch of words that is not useful in this research such as I, you, then. With stemming, words are reduced to their word stems [8]. For example, the sentence “He was swimming in the pool” will be reduced to “He is swim in the pool”.

C. TF-IDF

TF-IDF is an algorithm that is based on statistical values showing the appearance of a word in the document [9]. TF or Term-Frequency states how many words appear in a document. Meanwhile, IDF or Inverse Document Frequency states the number of documents that contain a word in one publication segment [6].

While computing TF, all words are treated equally important. However, it is known that certain words, such as “is”, “the”, and “and”, may appear a lot of times but have little importance. Therefore, we need to decrease the frequent terms while increasing the rare ones, by computing IDF, an inverse document frequency factor is integrated which depreciate the weight of words that occur very frequently in the document set and scale up the weight of words that occur rarely.

IDF is the inverse of the document frequency which measures the informativeness of word t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as “is” is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

Now there are few other problems with the IDF, in the case of a large corpus, the IDF value will explode. To avoid this effect, we use the log of the IDF. When a word that is not in vocab occurs, the DF will be 0. To avoid that, we add +1 to the denominator. Here is the final formula:

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \quad (1)$$

$$IDF(w) = log\left(\frac{N}{DF(w)}\right) \quad (2)$$

Where:

- *TF - IDF (w, d)*: weight of a word in all documents.
- *w*: a word
- *d*: a document
- *TF (w,d)*: the frequency of the occurring word *w* in document *d*
- *IDF (w)*: inverse *DF* from word *w*
- *N*: a total of document
- *DF (w)*: total document that has word *w*

D. Naive Bayes Classifier

Naive Bayes Classifier is a classification algorithm based on the Bayes theorem. Naive Bayes Classifier assumes that every word in a document is independent. That is the presence of one particular feature does not affect the other [4]. Bayes theorem formula is shown in (3).

$$P(B) = \frac{P(A)x\,P(A)}{P(B)} \quad (3)$$

E. Multinomial Naive Bayes

Multinomial Naive Bayes is very similar to Naive Bayes Classifier. Multinomial Naive Bayes calculate the multinomial distribution from each feature in the documents [10]. The probability of a document *d* being in class *c* is computed as:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \quad (4)$$

Where P(tk|c) is the conditional probability of term tk occurring in a document of class c. We interpret P(tk | c) as a measure of how much evidence tk contributes that c is the correct class. P(c) is the prior probability of a document occurring in class c. If a document's words do not provide clear evidence for one class versus another, we choose the one that has a higher prior probability. The formula of Multinomial Naïve Bayes is shown in (5).

$$Cmap = arg\ arg\ P(c) \prod_{k=1}^{m} P(t_k|c) \quad (5)$$

Parameter P(tk | c) (probability likelihood) is estimated by calculating the occurrence of tk on all training documents in c, using laplacean prior as shown in (6) [6]:

$$P(t_k \mid c) = \frac{1 + N_k}{|V| + N} \quad (6)$$

Where Nk counts the total occurrence of tk in c's training document and N is the total occurrence of words in c [6]. For example, given training this training data:

TABLE I. TRAIN DATA

Text	Class
Free money, click the link now	Spam
Where are you now?	Ham
I am busy now, call later	Ham
Click the link and get free souvenir from me	Spam

Then we want to classify the sentence "Call me for free item" where the target class is spam or ham. First, we will calculate the probability likelihood of every word in the sentence using (6).

TABLE II. PROBABILITY LIKELIHOOD

Word	P (word | spam)	P(word | ham)
Call	$\frac{0+1}{14+18}$	$\frac{1+1}{14+18}$
Me	$\frac{1+1}{14+18}$	$\frac{0+1}{14+18}$
For	$\frac{0+1}{14+18}$	$\frac{0+1}{14+18}$
Free	$\frac{2+1}{14+18}$	$\frac{0+1}{14+18}$
Item	$\frac{0+1}{14+18}$	$\frac{0+1}{14+18}$

Based on the calculation in Table 1, we can calculate the posterior probability of the sentence.

$$P(Call|spam) \times P(Me|spam) \times P(For|spam) \times P(Free|spam) \times P(Item|spam) = \frac{1}{32} \times \frac{2}{32} \times \frac{1}{32} \times \frac{3}{32} \times \frac{1}{32} = 0.000000178813934$$

$$P(Call|ham) \times P(Me|ham) \times P(For|ham) \times P(Free|ham) \times P(Item|ham) = \frac{2}{32} \times \frac{1}{32} \times \frac{1}{32} \times \frac{1}{32} \times \frac{1}{32} = 0.0000000596046448$$

Based on the result above, we can conclude that the sentence "Call me for free item" is classified as Spam because the sentence has a higher posterior probability in class Spam.

F. The Learning Process

This research uses TF-IDF along with Multinomial Naïve Bayes to classify a text as spam or ham. The learning process of the system can be seen in the following section.

1. First, we load the file that contains the list of text that want to be classified.
2. Then we begin the pre-processing that include tokenization, filtering, and stemming.
3. We also use stop-words to remove the list of words that occur too often.
4. When the pre-processing is over, we divide the dataset to train and test dataset.
5. And then we made a vocabulary of features as a training process.
6. After that, we convert the training dataset and test dataset to the TF-IDF model.
7. Finally, we fit the model to the Multinomial Naïve Bayes algorithm and then calculate the performance of the model.

G. Confusion Matrix

A confusion matrix is a method to measure the performance of the classifier. The confusion matrix is in the form of a table with 4 different combinations of predicted and actual values [11].

TABLE III. CONFUSION MATRIX

Predicted \ Actual	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Where TP occurred when the system predicted positive, and the actual data is positive (True Positive). TN means True Negative, where the system predicted negative, and the actual data is negative. FP occurred when the system predicted positive, but the actual data is negative (False Positive). FN means False Negative, where the system predicted negative, but the actual data is positive [11].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (8)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (9)$$

$$F1-Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (10)$$

Confusion matrix can be used to calculate precision, recall, F1-Score, and accuracy of the classifier. Accuracy is used to measure the accuracy of predictions made by the classifier. Precision calculates the ratio of true positive predictions to all positive predictions. While recall calculates the ratio of true positive predictions to all predictions in the actual class.

IV. RESULT AND DISCUSSION

A. Performance Evaluation

Several testing scenarios are used in this research to determine the best approach to classify spam feedbacks. The first scenario is to compare the performance of different ratios of train and test sets (Scenario 1). The next scenario is using upsampling and downsampling train dataset (Scenario 2). Upsampling is done by increasing the minority class data to the same amount as the majority class. Downsampling is done by reducing the majority class data to the same amount of minority classes. And then we compare both methods to determine what is the best method for our research. The last scenario is we use a typo-corrector from [12] in our research (Scenario 3). A typo-corrector is a method to fix a typo in the text. Our dataset has many words that contain a typo. Therefore, with the use of a typo-corrector, we believe it will increase the performance of the model.

In Scenario 1, we compare the performance from the 70:30 ratio of train and test set with 80:20. According to the result of both methods, each method obtained low precision, recall, and F1-Score. That is because the dataset used in this research is imbalanced. However, the best performance in this scenario is the ratio of 70:30 with 50% precision, 3% recall, and 6% F1-Score on spam class. The performance of this scenario can be seen in the table below.

TABLE IV. PERFORMANCE OF SCENARIO 1

Metrics	70:30		80:20	
	Spam	Ham	Spam	Ham
Precision	50%	89%	0%	91%
Recall	3%	100%	0%	99%
F1-Score	6%	94%	0%	95%

The best performance of this scenario has an accuracy of 94.3% in the training dataset and 92.59%

in the test dataset. The overall accuracy of this scenario can be seen in the table below.

TABLE V. ACCURACY OF SCENARIO 1

Metrics	70:30		80:20	
	Train	Test	Train	Test
Accuracy	94.3%	92.59%	90%	90%

Because our dataset is imbalanced, then our next scenario is applying an up-sampling and down-sampling method to our train data. With up-sampling and down-sampling, the performance of this imbalanced dataset is better than in the previous scenario. Up-sampling method got 46% precision, 44% recall, and 45% F1-Score. The down-sampling method got 25% precision, 67% recall, and 37% F1-Score on spam class. Up-sampling method has higher performance in precision and F1-Score, while the down-sampling method has higher scores in the recall. Therefore, it is necessary to decide which metric is a better value in this case.

In requirements engineering, user feedback is used by developers to find out what user needs of the application are used. Therefore, in this case, recall is a better metric in our research. However, when dealing with an imbalance dataset we can not only look at the value of precision or recall. Because our dataset is imbalanced, the recall value could not be considered accurate. We need to consider another value that is a harmonic value of both precision and recall, which is F1-Score. According to the F1-Score value and other overall scores, up-sampling is a better method in our research.

The performance and accuracy of scenario 2 can be seen in the table below.

TABLE VI. PERFORMANCE OF SCENARIO 2

Metrics	Up-sampling		Down-sampling	
	Spam	Ham	Spam	Ham
Precision	46%	94%	25%	94%
Recall	44%	93%	67%	78%
F1-Score	45%	94%	37%	81%

TABLE VII. ACCURACY OF SCENARIO 2

Metrics	Up-sampling		Down-sampling	
	Train	Test	Train	Test
Accuracy	98,67 %	88,88 %	98,41 %	77,03 %

In scenario 3, we apply typo-corrector to our dataset. In our dataset, there are many words used by users that are not listed in the dictionary and there are many typo errors. When we try to apply the typo-corrector to our dataset, it shows a better performance in our model. In this scenario, our model has the best performance of 89.25% accuracy, 45% precision, 56% recall, and 50% F1-Score on spam class as shown in Table 8. From this scenario, we could say that the current typo-corrector improves the performance of the model. However, the improvement is not significant. The result of this scenario could be seen in Table 8 and Table 9.

TABLE VIII. PERFORMANCE OF SCENARIO 3

Metrics	With Typo-corrector		Without Typo-corrector	
	Spam	Ham	Spam	Ham
Precision	45%	95%	45%	94%
Recall	56%	93%	48%	93%
F1-Score	50%	94%	46%	94%

TABLE IX. ACCURACY OF SCENARIO 3

Metrics	With Typo-corrector		Without Typo-corrector	
	Train	Test	Train	Test
Accuracy	98,49 %	89,25 %	98,67%	88,88 %

TABLE X. BEST MODEL PERFORMANCE

Metrics	Spam	Ham
Precision	45%	95%
Recall	56%	93%
F1-Score	50%	94%
Accuracy	89.25%	

Based on the test scenarios that have been done before, we conclude that using an upsampling method with the ratio of 70:30 in training and test data as well as the use of a typo-corrector succeeded in improving model performance that has an imbalanced dataset. The best model of these scenarios has 45% precision, 56% recall, 50% F1-Score, and 89.25% accuracy on spam class. For the ham class, this model has 95% precision, 93% recall, and 94% F1-Score.

V. CONCLUSION AND FUTURE WORKS

In this research, we used TF-IDF as feature extraction and Multinomial Naive Bayes to classify whether the user feedback of tiket.com application is spam or not. Our dataset consists of 810 ham and 90 spam data that are gathered using the scrapping method from the google play store. We use a confusion matrix to calculate the performance of the classification model. Based on previous test scenarios, the best model is reached when we apply upsampling method with the ratio of 70:30 in training and test data as well as using typo-corrector from [12]. The best performance of the model is 89.25% accuracy, 45% precision, 56% recall, and 50% F1-Score in the spam class.

In future works, we suggest adding more data to our current dataset because our dataset has very low spam data. We recommend that at least the ratio of minority and majority class is 40:60 so it will be balanced. Typo-corrector has been proven to improve our model. However, the improvement is still very low. An improvement to the typo-corrector could make a better performance in the classifier. Last, we suggest trying a different method of upsampling and downsampling, since that method is very useful when handling an imbalanced dataset. There are a couple of upsampling methods like SMOTE (Synthetic Minority Over-Sampling Technique) that could perform better than our previous upsampling method.

ACKNOWLEDGEMENT

We would like to express our special gratitude to the Software Engineering Laboratory in Universitas Multimedia Nusantara who supported us a lot in doing our research.

REFERENCES

[1] Wahono, R.S. (2006). *MENYEGARKAN KEMBALI PEMAHAMAN TENTANG REQUIREMENT ENGINEERING*. [Online] Available at: https://romisatriawahono.net/2006/04/29/menyegarkan-kembali-pemahaman-tentang-requirement-engineering/ [Accessed 1 Feb. 2020]

[2] Ceban, V. (2018). *User Feedback: What It Really Is and Why It's So Important to Any Business' Success?*. [Online] Available at: https://www.appzi.com/what-is-user-feedback-and-its-role/ [Accessed 1 Feb. 2020].

[3] Maxmonroe, n.d. *Arti Spam*. [Online] Available at: https://www.maxmanroe.com/vid/teknologi/arti-spam-adalah.html [Accessed 20 Sep. 2019]

[4] Gandhi, R. (2018). *Naive Bayes Classifier*. [Online] Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed 20 Sep. 2019]

[5] Matwin, S. & Sazonova, V. (2012). *Direct comparison between support vector machine and multinomial naive bayes algorithms for medical abstract classification*. [Online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422847/ [Accessed 1 Feb. 2020]

[6] Kalokasari, D. H., Shofi, D.I.M. & Setyaningrum, A.H. (2017). IMPLEMENTASI ALGORITMA MULTINOMIAL NAIVE BAYES CLASSIFIER PADA SISTEM KLASIFIKASI SURAT KELUAR. *JURNAL TEKNIK INFORMATIKA*, vol. 10, no. 2, pp. 109-118

[7] Arora, I. (2017). *Document feature extraction and classification*. [Online] Available at: https://towardsdatascience.com/document-feature-extraction-and-classification-53f0e813d2d3 [Accessed 18 May 2020]

[8] Heidenrich, H. (2018). *Stemming? Lemmatization? What?*. [Online] Available at: https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8 [Accessed 20 Sep. 2019]

[9] Wijaya, A.P. & Santoso, H.A. (2016). Naive Bayes Classification pada Klasifikasi Dokumen. *Journal of Applied Intelligent System*, vol. 1, no.1, pp. 48-55

[10] Huang, O. (2017). *Applying Multinomial Naive Bayes to NLP Problems: A Practical Explanation*. [Online] Available at: https://medium.com/syncedreview/applying-multinomial-naive-bayes-to-nlp-problems-a-practical-explanation-4f5271768ebf [Accessed 20 Sep. 2019]

[11] Narkhede, S. (2018). *Understanding Confusion Matrix*. [Online] Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 [Accessed 20 Sep. 2019]

[12] Setiabudi, Reza. (2020). Implementasi Algoritma Levenshtein Distance untuk Typo Correction Bahasa Indonesia pada User Feedback Aplikasi. Bachelor Thesis thesis, Universitas Multimedia Nusantara.

ISSN 2085-4552

Classification of Metagenome Fragments With Agglomerative Hierarchical Clustering

Alex Kurniadi[1], Marlinda Vasty Overbeek[2]
Informatics Study Program, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1]alex.kurniadi@student.umn.ac.id, [2]marlinda.vasty@umn.ac.id

Accepted 8 July 2021
Approved 12 August 2021

***Abstract*—Unlike genomics which study specifically culturable microorganisms, metagenomics is a field that studies microorganic samples retrieved directly from the environment. Such samples produce widely varying fragments when sequenced, many of which are still unidentified or unknown. Assembly of these fragments in the goals of identifying the species contained among them are thus prone to make said goals more difficult, so it becomes necessary for binning techniques to come in handy while trying to classify these mixed fragments onto certain levels in the phylogenetic tree. This research attempts to implement algorithms and methods such as k-mers to use for feature extraction, linear discriminant analysis (LDA) for dimensionality reduction, and agglomerative hierarchical clustering (AGNES) for taxonomic classification to the genus level. Experimentation is done across different objective measurements, including the length of the observed metagenome fragment that spans from 0.5 Kbp up to 10 Kbp for both the 3-mer and 4-mer contexts (k = 3 and k = 4). The averaged validity scores of the resulting data clusters generated from both the training and test sets, computed with the silhouette index metric, are 0.6945 and 0.0879 for the 3-mer context, along with 0.5219 and 0.1884 for the 4-mer context.**

***Index Terms*—AGNES; k-fold; k-mers; LDA; machine learning; metagenomics**

I. Introduction

In an effort to analyze the genetic material of nonculturable microorganisms (which cannot be cultured in laboratories), samples were taken directly from the environment. The sample taken may contain fragments of genetic material (genome) from a variety of different species. When the sequencing and assembly procedures are carried out on this mixture of fragments simultaneously, the mismatch between the genomes of one species with another will result in chimeric contigs that lead to the phenomenon of interspecies chimerae, so that the species diversity of the sample cannot be known [1] [2]. The term "contig" itself is taken from the English word "contiguous", and is defined as a strand of genomic fragments (DNA) of a species that are close together, representing a subset of DNA [3]. Chimeric contigs are then defined as a single contig strand composed of genomic fragments from two or more different species [4]. To minimize the chance of these occurrences, it is necessary to apply a binning technique so that each distinctive fragment of the compound can be separated as well as possible from one another.

This binning technique has two approaches, namely homology-based binning and composition-based binning. The homology approach was carried out by aligning the sample metagenome fragment sequences against the sequence data from the NCBI and concluding the results at the taxonomic level. Meanwhile, the compositional approach uses the result of feature extraction in the form of base pairs as input for the learning model [2].

There are two main approaches of learning for machine learning models, namely learning by example (supervised learning) and learning by observation (unsupervised learning). In the context of classification, supervised learning already has categorization information in its learning base, while unsupervised learning only has training data as a learning base. The method used in this study belongs in the realm of unsupervised learning.

In this study, metagenome fragments in the data from NCBI sources will be grouped using the *k*-mers method for feature extraction, linear discriminant analysis (LDA) method for data dimension reduction, and agglomerative hierarchical clustering (AGNES) algorithm for grouping. The *k*-mers method was chosen to be used in this study because this method works by calculating the number of occurrences of short strands (substring; polymer) along *k* letters in one genome strand, which will later highlight characteristic distinctions based on differences in the number of frequencies of each polymer between individual samples [5]. The LDA method was chosen because this method aims to try to explain a dependent variable as output (genus taxonomic level of the studied data) based on the values of the independent variables as input (genome strands belonging to the sample data). Meanwhile, the agglomerative hierarchical clustering method was chosen on the basis of its bottom-up workflow, because metagenomic fragment analysis starts from base pair units which then forms various long strands from each fragment based on the frequency of occurrence of certain base pair combinations.

II. Literature Study

A. Metagenomics

Metagenomics is a branch of science that studies genetic material in samples taken directly from the environment [2]. Unlike genomics, where the analysis is carried out only on certain isolated (or previously cultured) organisms, metagenomic analysis is carried out directly on a group of microorganism communities without the need for first culturing efforts. This has brought interesting insights into the ecological systems of various habitats [6]. The emergence of this research field was triggered by the development of sequencing technology, with Roche's 454 pyrosequencing technique as an example [7].

B. Machine Learning

Machine learning is a field of science that studies computer algorithms that develop independently through experience [11]. Models built for machine learning work by taking input in the form of training data as learning material to analyze new, foreign data (test data). The learning approaches that are commonly applied in designing machine learning models consist of supervised learning and unsupervised learning. Supervised learning refers to a learning process in which the training data contains relevant input and output information, while in unsupervised learning, the training data only contains input information, so the model must draw its own conclusions (outputs) based on the learned input [8].

C. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a dimensionality reduction method that selects and extracts a set of the most discriminatory features from the data for multi-class classification [9]. This method is regarded as one of the most useful and popular methods with various applications, belonging in the realm of clustering algorithms [10]. The main difference between LDA and another dimensionality reduction method that is also frequently used, principal component analysis (PCA), lies in the main focus of each method. LDA and PCA aim to find the component with the highest variance in the data, but LDA also prioritizes the level of separability between data classes [11].

The LDA method can be summarized into seven steps as follows [11].

1) Standardize the initial d-dimensional data, where d represents the amount of features present in the data.

2) Compute the d-dimensional mean vectors m_i for each class in the data.

$$m_i = \frac{1}{n_i}\sum_{x \in D_i} x_m \tag{1}$$

$$m_i = \begin{bmatrix} \mu_{i(\text{feature } 1)} \\ \mu_{i(\text{feature } 2)} \\ \vdots \\ \mu_{i(\text{feature } n)} \end{bmatrix}^T, i \in \{1, 2, \dots, c\} \tag{2}$$

3) Compute the between-class scatter matrix S_B and the within-class scatter matrix S_W.

$$S_W = \sum_{i=1}^{c} S_i \tag{3}$$

$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^T \tag{4}$$

$$S_B = \sum_{i=1}^{c} n_i\,(m_i - m)(m_i - m)^T \tag{5}$$

4) Compute the eigenvectors along with the respective eigenvalues from the matrix $S_W^{-1}S_B$.

5) Sort the computed eigenvalues in descending order.

6) Take k eigenvectors with the highest eigenvalues to form the $d \times k$-dimensional transformation matrix W, where each eigenvector acts as one column.

7) Use the transformation matrix W to transform the initial d-dimensional data matrix X into a new feature matrix of dimension k.

Feature data that has been extracted is first normalized with min-max scaling method into the range [0, 1]. This normalization is applied to each feature with the formula below [11]:

$$x_{normal} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6}$$

For each feature, the normalized value x_{normal} is obtained by calculating the weight of the individual values of x against the range of values in that feature, namely the difference between the largest and smallest values in said feature, denoted as x_{max} and x_{min}.

After computing the within-class and between-class distribution matrices, the eigenvectors and eigenvalues can be obtained by solving the matrix $S_W^{-1}S_B$. With the largest eigenvalues obtained, the corresponding eigenvectors are then combined into columns for the transformation matrix W. The transformation process is then carried out by multiplying the matrix W as shown in (7), where X is the initial data matrix and X' the new, reduced feature matrix:

$$X' = XW \tag{7}$$

D. Hierarchical Clustering

Hierarchical clustering is a method in statistics for performing clustering. This method works by first

ISSN 2085-4552

measuring the level of inequality between two clusters of data before combining the two clusters into a new cluster (agglomerative) or splitting each cluster into two new clusters (divisive). The advantages of this method include ease of understanding and application and the absence of an obligation to first know the number of clusters desired [12]. The difference between agglomerative and divisive methods lies in the flow of work, where agglomerative hierarchical clustering forms large clusters of data by combining individual samples or a collection of small clusters (bottom-up), while divisive hierarchical clustering forms small clusters of data by splitting up clusters of data. larger [12]. The hierarchical clustering method that will be used in this study is the agglomerative method.

The level of inequality is measured based on the distance metric used between two data clusters, and the merging or splitting of the two clusters is carried out based on the linkage criterion used. There are a number of metrics and criteria commonly used in hierarchical clustering research, including the Euclidean distance metric (8) and the single-linkage linkage criterion (abbreviated as SLINK; (9)) [13]:

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{8}$$

$$\min\{d(a,b) \mid a \in A, b \in B\} \tag{9}$$

E. Silhouette Index

The performance measurement of models using clustering algorithms is different from the measurements on classification algorithms in general. In evaluating clustering performance, the measured qualities are the closeness between samples in the same cluster, the separation of one sample from similar samples that are part of different clusters, and the separation between clusters [14]. To measure the validity of the results of such an algorithm, there are several different benchmark systems that can be used, and one such system used in this study is the silhouette index.

Silhouette index works by comparing the average distance of an individual sample *i* to all other samples in the same cluster, C_i, with the closest distance of the sample to all other clusters, C_k. The formula for the silhouette index is as follows.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i), b(i)\}}, if\ |C_i| > 1 \tag{10}$$

The variable *a(i)* represents the average distance of an individual sample *i* to every other sample *j* in the C_i group, *b(i)* represents the distance of a sample *i* to the nearest other cluster C_k, and *s(i)* states the silhouette index value for the sample *i*. In the case where a C_i cluster only has *i* as its sole sample, then *s(i)* will be zero. The formulas for two sample distances *a(i)* and *b(i)* can be seen in (11) and (12).

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{11}$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \tag{12}$$

III. Methodology

A. Preparing the Genome Data

Genomic data for each observed microorganism species is obtained from the NCBI website and processed using the MetaSim tool to generate metagenome data in FASTA format files. The data set used in this study consists of eighty species from ten genera. Research will be carried out on different fragment lengths, ranging from 0.5 Kbp (kilo base pair; 10^3 bp), 1 Kbp, 5 Kbp, up to 10 Kbp. The total weighting of all species in the final fragment is applied to 10,000.

B. Feature Extraction

The resultant FASTA data is first preprocessed before the data can be used by the machine learning model. The preprocessing procedure here begins with feature extraction using the *k*-mers method, where each strand of fragments is observed to note how often each combination of base pairs A, T, G, and C, as a key component determining the characteristics of microorganisms, appears on the thread. The length of base pair combinations that will be observed here are 3-mers such as AAA, AAC, AAG, up to TTC, TTG, TTT (4^3 = 64 combinations), and 4-mers such as AAAA, AAAC, up to TTTG, TTTT (4^4 = 256 combinations), as shown in Fig. 1. This distinction between combinations is important, because DNA is the genetic material in almost all living things, including microorganisms and humans, where different combinations will yield different physiological characteristics [16].

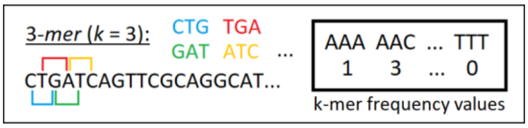


Fig. 1. Illustration of 3-mer frequency counting

C. Data Normalization

The extracted features data is then normalized using the min-max scaling method. In min-max normalization, all feature values are rescaled into a new range of values while maintaining the weight of each value. In this study, the range of values used is [0, 1], as can be seen in Fig. 2.

AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
10	7	0	3		5	5	4	9
2	8	9	1		6	11	0	3
3	10	6	0		17	4	2	16
7	1	4	3		4	5	6	3

AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
0.588235	0.411765	0	0.176471		0.294118	0.294118	0.235294	0.529412
0.117647	0.470588	0.529412	0.058824		0.352941	0.647059	0	0.176471
0.176471	0.588235	0.352941	0		1	0.235294	0.117647	0.941176
0.411765	0.058824	0.235294	0.176471		0.235294	0.294118	0.352941	0.176471

Figure 2. Illustration of min-max scaling

D. *Dimensionality Reduction*

After normalization is complete, dimensionality reduction is carried out on the data set before it is ready to be used in machine learning models. Dimensionality reduction aims to eliminate redundant characteristics by transforming the characteristics from the data matrix with larger dimensions into smaller dimensions [16]. The reduction algorithm used in this study is Linear Discriminant Analysis (LDA). In this study, the LDA algorithm is run by following the steps described in [11], except for data standardization. This is because the data to be reduced has already been normalized, so the LDA procedure here can be continued without having to standardize the data again.

E. *Clustering*

Analysis is then carried out on the training data set first to see and determine the most optimal conditions. The optimal conditions are to be used as the basis for the analysis of the test data set in order to group the data set into genera. The algorithm used for this testing phase is agglomerative hierarchical clustering.

To measure the validity of each data cluster resulting from the model grouping, there are several benchmark systems that can be used, including the silhouette index used in this study. Silhouette index of each sample is calculated based on the average distance of the sample to all other samples in the same cluster and its distance to all other clusters by taking the cluster closest to the sample.

IV. DISCUSSION

A. *Metagenome Data*

For the sample data, metagenome DNA fragment strands processed by MetaSim were prepared from 80 species of microorganisms belonging to 10 different genera. The total weighting of all species for the metagenome fragments is 10,000.

B. *Feature Extraction*

The preprocessing stage begins by performing feature extraction in the form of base pair combinations from the metagenome fragment data. After the feature extraction is successful, a new set of numerical data is obtained which can later be used by the machine learning model. The number of data rows in the sample table corresponds to the results of the previous MetaSim processing, while the number of data columns adjusts to the observed *k*-mer value, which is 4^k columns.

C. *Normalization*

The sample data is first normalized using the min-max scaler from the preprocessing module scikit-learn [14]. The values in each column are calculated and converted to the relative weights of each value against all other values in the same column. The range of values used as a reference for normalization is [0, 1], which means that the lowest value in the column becomes 0, while the highest value in the same column becomes 1.

D. *Dimensionality Reduction*

After normalizing the data, dimensionality reduction is performed to obtain a new, smaller matrix containing the projection of the sample data set matrix while maintaining the integrity of information between data classes. This stage begins by calculating the mean vector for each existing class, resulting in a collection of *d*-dimensional vectors where *d* is the number of features.

With the mean vectors, the within-class and between-class distribution matrices S_W and S_B are computed. These two distribution matrices are then used to obtain a series of eigenvectors and eigenvalues as indicators of the integrity of information between classes in the data. Eigenvalues that are not close to zero are taken to create a transformation matrix *W*, which will later be used to transform the $n \times d$-dimensional sample data set into an $n \times k$-dimensional matrix, where *k* does not exceed the number of features (*d*) nor the number of classes subtracted by one (c - 1).

E. *Data Splitting*

The sample data set is divided into training data and test data, with the proportion of test data being 20% of the initial set.

F. *Clustering*

The reduced data set is then grouped into clusters by the model, using agglomerative hierarchical clustering (AGNES) algorithm based on information from the training data. Because this model aims to group the test data into genera that have been presented previously, the number of clusters adjusts to the number of existing genera, which is 10 clusters. The distance metric used is Euclidean, and the relationship criterion used is single-linkage, where the shortest distance between two data points from two different clusters is taken as the distance between the two clusters.

G. *Evaluation*

Evaluation of the machine learning model is carried out using the silhouette index assessment method, in which each individual sample processed by the model is assessed based on the sample's proximity to its own cluster with the closest other cluster [14]. The final value of the silhouette index of the learning model is obtained by averaging values of the silhouette index for the entire data set. The range of values in this scoring method is [-1, 1], with -1 as the worst value and 1 as the best value. A value that tends to be negative indicates that the samples are grouped into the wrong clusters, whereas a value close to 0 (zero) indicates that the clusters tend to overlap with each other [14].

The process of evaluating the model's performance in grouping sample data spans across a series of observational contexts on both training (80% sample size) and test (20% sample size) data sets. These contexts consist of the variations in total length of metagenome fragments (0.5 Kbp to 10 Kbp) and the polymer length (3-mer and 4-mer), as shown in Tables I and II below.

TABLE I. TRAINING DATA SILHOUETTE INDEXES

Total fragment length	***k*-mer**	
	***k* = 3**	***k* = 4**
0.5 Kbp	0.6022	0.0162
1 Kbp	0.6285	0.1318
5 Kbp	0.7443	0.0403
10 Kbp	0.8029	0.1634
Average	**0.6945**	**0.0879**

TABLE II. TEST DATA SILHOUETTE INDEXES

Total fragment length	***k*-mer**	
	***k* = 3**	***k* = 4**
0.5 Kbp	0.4397	0.1083
1 Kbp	0.4678	0.0572
5 Kbp	0.5444	0.2160
10 Kbp	0.6356	0.3724
Average	**0.5219**	**0.1885**

From the four average values above, it is then known that the data grouping quality of the model in the 3-mer context is better than that in the 4-mer context (0.6945 > 0.0879; 0.5219 > 0.1885). In the 3-mer context, the silhouette index value being closer to 1 indicates that the data has been grouped fairly well and is quite separate between each data cluster. Meanwhile, in the 4-mer context, the silhouette index being closer to 0 indicates that after data is grouped by the model, there are many data clusters that overlap with each other, so that the separation between data clusters becomes unclear.

V. CONCLUSION

In this study, the metagenomic fragment data set was first preprocessed with LDA as the dimensionality reduction method and *k*-mers as the feature extraction method. The preprocessed data set was then grouped with agglomerative hierarchical clustering algorithm and the resulting clusters were evaluated with the silhouette index metric.

The silhouette index values, a measure of the validity of data grouping by the model, ranged from 0.6022 ~ 0.8029 for the training set and 0.4397 ~ 0.6356 for the test set in the 3-mer context. In the 4-mer context, the silhouette index values ranged from 0.0162 ~ 0.1634 for the training set and 0.0572 ~ 0.3724 for the test set. This means that in the 3-mer context, the resulting data have been clustered quite well and the clusters are quite separate between each other. However, in the 4-mer context, there are many clusters that overlap with each other, causing the silhouette index value to come close to zero.

REFERENCES

[1] Overbeek, M. V., Kusuma, W. A. and Buono, A. (2013) 'Clustering metagenome fragments using growing self organizing map', *2013 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2013*, pp. 285–289. doi: 10.1109/ICACSIS.2013.6761590.

[2] Simangunsong, V. F. R. (2015) 'Klasifikasi Fragmen Metagenom Menggunakan Principal Component Analysis Dan K-Nearest Neighbor', pp. 1–34.

[3] Gregory, S. G. (2005) 'Contig Assembly', *Encyclopedia of Life Sciences*, pp. 1–4. doi: 10.1038/npg.els.0005365.

[4] Scholz, M. *et al.* (2020) 'Large scale genome reconstructions illuminate Wolbachia evolution', *Nature Communications*, 11(1). doi: 10.1038/s41467-020-19016-0.

[5] Rosen, G. *et al.* (2008) 'Metagenome Fragment Classification Using -Mer Frequency Profiles', *Advances in Bioinformatics*, 2008, pp. 1–12. doi: 10.1155/2008/205969.

[6] Richter, D. C. *et al.* (2008) 'MetaSim - A sequencing simulator for genomics and metagenomics', *PLoS ONE*, 3(10). doi: 10.1371/journal.pone.0003373.

[7] Margulies, M. *et al.* (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

[8] Bishop, C. M. (2006) Pattern Recognition and Machine Learning, The Ecstatic and the Archaic: An Analytical Psychological Inquiry. Springer.

[9] Yan, C. *et al.* (2020) 'Self-Weighted Robust LDA for Multiclass Classification with Edge Classes'. Available at: http://arxiv.org/abs/2009.12362.

[10] Liu, J. *et al.* (2020) 'Capped norm linear discriminant analysis and its applications', (1), pp. 1–11. Available at: http://arxiv.org/abs/2011.02147.

[11] Raschka, S. and Mirjalili, V. (2019) *Python Machine Learning*. Third. Birmingham: Packt Publishing Ltd.

[12] Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. 99th edn. Hoboken: John Wiley & Sons, Inc.

[13] SAS Institute Inc. (2009) *SAS/STAT(R) 9.2 User's Guide, Second Edition*. 2nd edn. Edited by A. Jones and E. Huddleston. Cary, NC: SAS Institute Inc.

[14] Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: http://scikit-learn.sourceforge.net. (Accessed: 3 June 2021).

[15] *What is DNA?: MedlinePlus Genetics* (2021). Available at: https://medlineplus.gov/genetics/understanding/basics/dna/ (Accessed: 28 June 2021).

[16] Tharwat, A. *et al.* (2017) 'Linear discriminant analysis: A detailed tutorial', *AI Communications*, 30(2), pp. 169–190. doi: 10.3233/AIC-170729.

ISSN 2085-4552

Elastic Stack Ability Test Monitoring Slowloris Attack on Digital Ocean Server

Is Mardianto[1], Dedy Sugiarto[2], Krisna Aditama Ashari[3]
Informatics Engineering, Faculty of Industrial Technology, Universitas Trisakti, Jakarta, Indonesia
[1]mardianto@trisakti.ac.id, [2]dedy@trisakti.ac.id, [3]krisna064001904003@std.trisakti.ac.id

Accepted 31 July 2021
Approved 12 September 2021

***Abstract*— Servers have a central role in computer network. The server is in charge of serving user requests with various types of services. Every server activity in handling these things will generate different types of logs. Information from this large amount of logs is often ignored and has not been widely used as material for analyzing the performance of the server itself. In this study, Elastic Stack is functioned as a system that handles upstream to downstream processes starting from collection, transformation, and storage as well as graphical visualization of the Nginx web server given an attack scenario in the form of massive incoming connection requests and server login access attempts. The Elastic Stack components used as log collectors are Filebeat and Metricbeat for system metric data. For testing attacks using the Slowloris tool which will consume web server resources. The results of the research that have been carried out are when there are 500 incoming connections, the web server can serve requests normally, at 1000 connections there are some packets that are not served, the server becomes unable to access when it reaches a total of 2000 incoming connections. Metric data in the form of CPU Usage and Memory Usage are affected, although not significantly. Identification of IP Address shows the source of the attack comes from Singapore, according to the domicile of the attacker's computer. All access data in the form of username, time, origin of region trying to enter the server are recorded by the system.**

***Index Terms*—availability; cloud computing; filebeat; log; metricbeat.**

I. Introduction

Log is a file containing a list of events that occur on a computer system [1]. This log file is also owned on a server that runs various types of services on it. As a system administrator, log files on these servers can provide useful information in monitoring server performance. CPU performance, memory usage, disk, network I/O, as well as the ability to detect disturbances from inside and outside the system are some important indicators in the availabilty or sustainability of server performance in running its services. Thus, a centralized logging system is needed to be able to transmit existing system status data for later reporting that is easy to analyze [2]. Some of the capabilities that today's modern log management systems must have include: Aggregation, namely the ability to collect and transmit logs from various data sources. Processing to convert log messages into meaningful data. Data storage for a long time to allow monitoring, trend analysis, and security, as well as Analysis, which is to sort data by performing queries and making visualizations [3].

Elastic Stack or previously known as ELK Stack is a collection of open source software developed by Elastic which is useful for searching, analyzing, and visualizing logs generated from any source in any format [4].

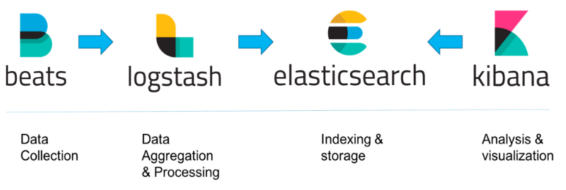


Fig. 1. Elastic Stack Components

The main components of the Elastic Stack are divided into 4 parts as shown in Fig. 1, Elasticsearch is a place where log data is stored and indexed, Logstash functions to process various sources and types of data into more structured information, and Kibana which is designed as a visualization platform, providing a web-based interface to search, view, and analyze data stored in Elasticsearch clusters [5].

Another important component is Beats. Beats is a platform that handles sending data from various sources [6]. Filebeat and Metricbeat are part of Beats. Filebeat is a data transmission platform that collects and transmits data from various sources and then forwards the data to the Elastic Stack [7]. Filebeat can be installed on almost any operating system, including as a Docker container, and it also comes with built-in modules for certain platforms (such as Apache, MySQL, Docker, MariaDB, Percona, Kafka, etc.). Metricbeat will periodically pass metric data from the operating system and service statistics running on the system to Elasticsearch or Logstash [8].

The use of Elastic Stack on the server is the right solution in log management because it is able to display information to administrators about statistics, trends, and anomalies that arise [9]. These are evidenced in several previous studies regarding the use of the Elastic Stack, including research on monitoring the apache web server in handling user requests [1], generating graphs of web server performance against time, utilization and error, other studies also collect logs from campus network infrastructure from various devices [2], generates a collection of log data by category such as warnings, errors, information and notifications. In addition, Elastic Stack can also be used as a detection system and information collection related to attempts to access the server and delete log data by the client [10]. Related research to analyze the types of slow HTTP attacks and their impact on virtual machines using the DSTAT tool, the results obtained indicate an increase in server resource consumption such as CPU and memory usage [11]. The difference between this research and previous research is in the use of Digital Ocean's cloud computing servers, as well as measurement and analysis of server performance in running the Nginx service against Slowloris attacks using Elastic Stack.

Nginx is an open-source web server software, created by Igor Sysoev and released to the public in October 2004. At its release, the makers assured the public that Nginx can solve web server performance problems in handling large numbers of active connections simultaneously. Nginx provides lower memory usage than other web servers, and also has the following features: reverse proxy, IPv6, load balancing, FastCGI support, web sockets, static file processing, TLS/SSL [12].

Slowloris is one such attack tool by opening a connection, then sending HTTP headers, adding them but never completing the request. Thousands of HTTP POST connections are made and HTTP headers are sent very slowly to force the target web server to keep the connection open. Slowloris will take all the resources of the target web server, thereby blocking requests from legitimate clients or clients who want to access the web server [13]. This attack belongs to the category of availability attacks where the server is not available when needed.

How to measure the impact on the system, and Nginx service when experiencing a Slowloris attack and analyze the source early is the purpose of this study. Therefore, this study aims to use the Elastic Stack as a system that helps collect and process various system log data with examples of Slowloris attacks.

II. METHOD

The method used in this study consists of several stages as shown in Fig. 2.

A. Preparation

This stage is in the form of designing and preparing a cloud server that will function as a log server and web server. The infrastructure uses the services of a cloud computing service provider, namely DigitalOcean. The operating system used is Ubuntu version 20.04. The amount of cloud computing costs varies depending on the required server specifications, in this study using server specifications with details as shown in Table I.

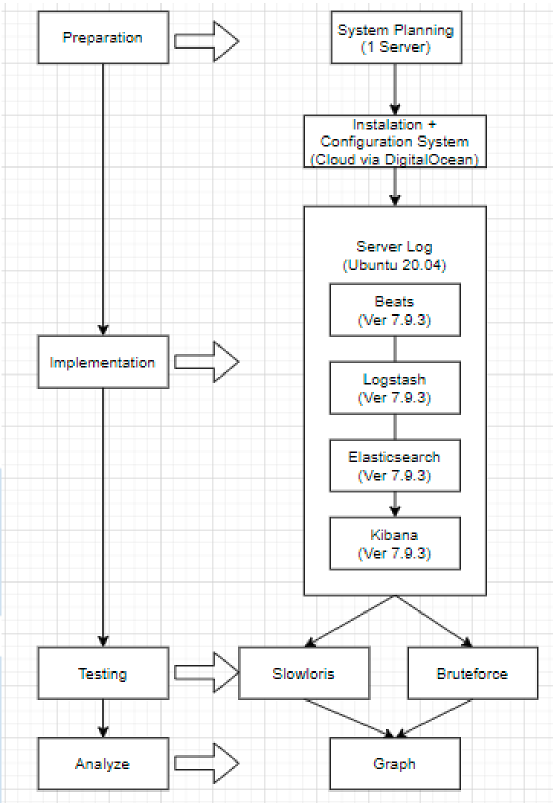


Fig. 2. Methodology

TABLE I. SERVER SPECIFICATIONS

No	Component	Description
1	Processor	2 CPU
2	RAM	4 GB
3	HDD	80 GB
4	Transfer	4 TB
5	Location	Singapore

After the process of selecting specifications and filling in data about the created server has been completed, the server is ready for use as shown in Fig. 3

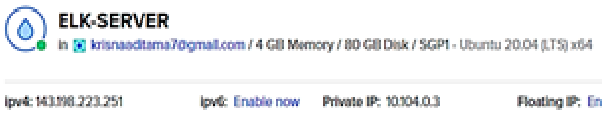


Fig. 3. Server on DigitalOcean

To perform an attack simulation, an additional PC is needed that functions to run Slowloris. The 2 additional PCs use the infrastructure of Google Cloud Platform as shown in Table II.

TABLE II. ATTACKER SERVER

No	Name	Zone	External IP

1	krs1	asia-southeast1-a	35.240.175.118
2	krs2	asia-southeast2-a	34.101.189.102

B. Implementation

The second stage includes the implementation and configuration of the Elastic Stack packages, namely Beats, Logstash, Elasticsearch and Kibana as the system that will manage the logs. And the use of Nginx as a web server.

1) Elastic Stack and Nginx

Because the server uses a cloud system, interaction with the server is carried out remotely through the PuTTY application using the Public IP that has been provided by the cloud provider, as well as the password that was previously set when creating the server. To be able to use the Elastic Stack on the server, it requires the java package to be pre-installed by adding the repository from the Elastic Stack web (https://artifacts.elastic.co).

The first package to create is Elasticsearch. Elasticsearch has the ability to index logs and can instantly search for specific records from billions of log records [14]. In Elasticsearch the configuration is done in the /etc/elasticsearch/elasticsearch.yml file, remove the hash mark or uncomment the network.host and http.port lines, set the ip on network.host to 0.0.0.0 and because it only runs on single server then add the line discovery.type: single-node. The next package needed is Logstash, Logstash acts as a data collector, forwarder, and processes log data into JSON format [15], so that the log data from filebeat can be processed later, it is necessary to create a configuration file. The logstash.conf file is stored in the /etc/logstash/conf.d/ folder. The logstash configuration file is saved under the name nginx.conf:

```
nginx.conf
Input: beats
Output: index

  input {
    beats {
      port => 5044
    }
  }
  output {
    if [@metadata][pipeline] {
      elasticsearch {
        hosts => localhost
        manage_template => false
        index => "%{[@metadata][beat]}-
        %{[@metadata][version]}-
        %{+YYYY.MM.dd}"
        pipeline                              =>
"%{[@metadata][pipeline]}"
      }
    } else {
      elasticsearch {
        hosts => localhost
        manage_template => false
        index => "%{[@metadata][beat]}-
        %{[@metadata][version]}-
        %{+YYYY.MM.dd}"
      }
    }
  }
```

In the configuration file, Logstash processes log input from Beats and produces an output in the form of an index per day based on the type of log data provided by Beats to be saved to Elasticsearch. To run the configuration use sudo /usr/share/logstash/bin/logstash -f /etc/logstash/conf.d/nginx.conf.

The last Elastic Stack package is Kibana. Kibana is an open-source tool for visualizing Elasticsearch data. Present users with various types of visual dashboards, such as bar charts, pie charts, time charts, histograms, heat maps, map visualizations, etc [16]. The configuration file that has been changed is /etc/kibana/kibana.yml by removing the hash mark / uncomment on the server.port: 5601 and server.host: 0.0.0.0 lines. After all packages are installed and configured, the service is run with the command sudo service service-name start. The service-name is changed according to the package name (elasticsearch / logstash / kibana), because the web server uses Nginx, use the default configuration for the service.

2) Beats

To be able to get log data it is necessary to install Beats. Beats is an agent for single-purpose data shippers. It sends various types of data (such as metrics, files, and networks) to Logstash [17]. To get the log collection functions of Nginx and the system running, add the system and nginx modules with the command sudo filebeat modules enable nginx system. Next run the command filebeat setup --pipelines --modules nginx,system. This module features an embedded fileset such as a source directory of logs to be collected, an ingest node pipeline for parsing data as well as providing a sample Kibana dashboard for graphical visualization of several frequently processed log formats. In the process of parsing the data through the ingest node of the Elasticsearch pipeline, transformations are made from the log lines such as 180.248.120.169 - - [03/Jul/2021:01:45:56 +0000] "GET /krisna HTTP/1.1" 404 197 " -" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.101 Safari/537.36" into a more structured form of data in JSON format.

In this ingest pipeline there is also a process of enriching information from the logs obtained with the help of the processor inside. Examples of processors used are GeoIP and User agent, where GeoIP provides detailed information regarding the country of origin, city and coordinates of the location of the incoming request, while the User agent will provide information related to the browser and operating system used by the user. Log the JSON form with additional information such as the following:

```
JSON Format Log

  "nginx": {
        "access": {
          "remote_ip_list": [
            "180.248.228.125"
```

```
        ]
      }
    },
    "log": {
      "file": {
        "path": "/var/log/nginx/access.log"
      },
      "offset": 563087
    },
    "source": {
      "geo": {
        "continent_name": "Asia",
        "region_iso_code": "ID-KI",
        "city_name": "Balikpapan",
        "country_iso_code": "ID",
        "region_name": "East Kalimantan",
        "location": {
          "lon": 116.8428,
          "lat": -1.2551
        }
      },
      "as": {
        "number": 7713,
        "organization": {
          "name": "PT Telekomunikasi
Indonesia"
        }
    }
```


In the /etc/filebeat/filebeat.yml file, uncomment the output.logstash and hosts: ["localhost:5044"] lines so that the logs obtained will be forwarded to Logstash for further processing.

To install the metricbeat package, use sudo apt-get install metricbeat. By default, metricbeat will collect system data such as CPU performance, memory usage, network traffic and other data on the system module. Finally, load all the configurations that have been made earlier with the command filebeat setup and metricbeat setup.

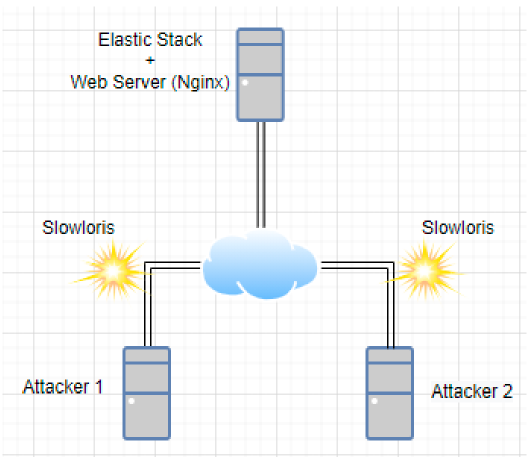


Fig. 4. Slowloris Attack Scheme

C. Testing

This testing phase includes testing the web server's ability to handle the maximum request load and can track the origin of the attack and record server login access attempts from outside parties. In the experiment, 2 PC servers were prepared as a source of attack that were installed using a cloud computing system, an illustration as shown in Fig. 4.

The test aims to obtain data on the impact of attacks recorded on the Elastic Stack, when several attack scenarios are carried out on the server as shown in Table III.

TABLE III. ATTACK SIMULATION

Scenario	Number of PCs	Attack (/PC)	Duration
1	1	500 Connection	20 Minutes
2	2	500 Connection	20 Minutes
3	2	1000 Connection	20 Minutes

In launching an attack on the web server, the Slowloris application was used which was obtained from the GitHub link, at the address https://github.com/gkbrk/slowloris.git, the attacker PC uses Git and clones the Slowloris file from the GitHub address above using the command git clone https://github.com/gkbrk/slowloris.git

To run Slowloris the command used is python3 slowloris.py <<Domain / IP Target>> -p 80 -s <<Number of Connections>> --sleeptime <<Attack lag time>>

D. Analysis

The last stage is viewing the logs that have been processed into graphic form, to access the Kibana service on the browser by opening the address http://IP-Public:5601. Various types of visualizations have been provided by the Beats modules, this certainly makes it easier to analyze server performance and identify problems.

III. RESULT AND DISCUSSION

A. Attack Impact Testing

In the testing scenario 1 as shown in Fig. 5, the web server is loaded with request traffic of 500 connections originating from 1 PC using Slowloris for 20 minutes.

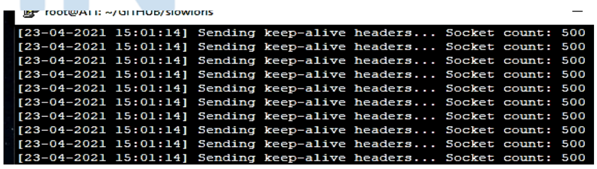


Fig. 5. Slowloris attack 500 connections from 1 PC

ISSN 2085-4552

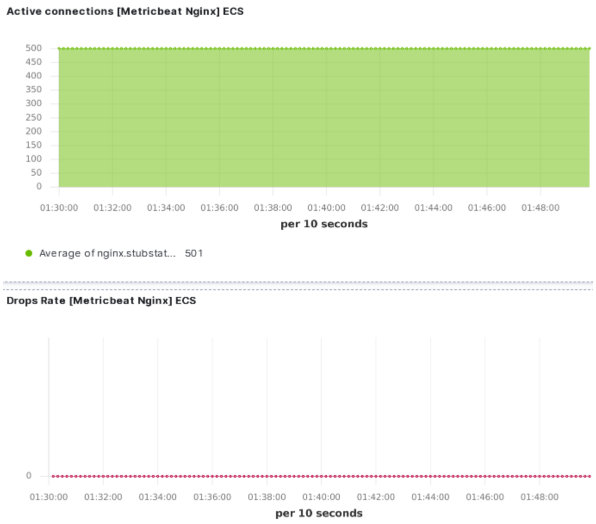


Fig. 6. Web Server Log Results in Scenario 1

Fig. 6 is the Nginx status chart in Kibana, show that the service is running normally where all 500 incoming connections can be served with a stable drop rate indicator at 0.

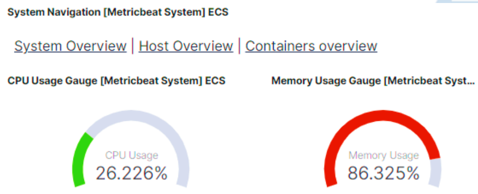


Fig. 7. CPU Usage and Memory Usage in Scenario 1

The results of the log observations as show on Fig. 7, CPU Usage indicator recorded during the attack was 26.226% and Memory Usage 86.325%. From the data obtained, it shows that the web server resource is still sufficient to serve all incoming requests. This is because the number of incoming attacks is still below the limits of the CPU and memory capabilities in handling incoming processes, so the attacks are not yet at the stage of disrupting the running of the computer system.

In testing scenario 2 as shown in Fig. 8, the web server is loaded with request traffic of 500 connections originating from 2 PCs using Slowloris with the same duration of 20 minutes.

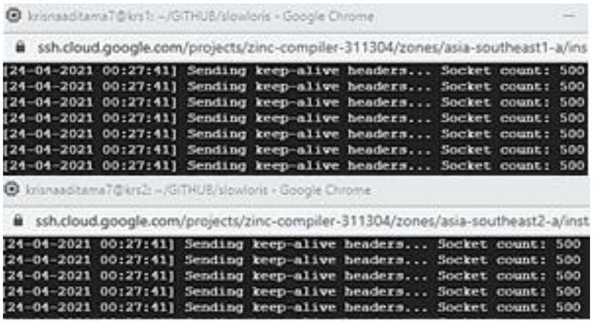


Fig. 8. Slowloris attack 500 connections from 2 PCs

Fig. 9. Web Server Log Results in Scenario 2

The results of observations as shown in Fig. 9 show that there were several times the server failed to handle requests, with various drop rate indicators (100, 400, 700).

In the information shown in Fig. 10, CPU Usage and Memory Usage has increased compared to the first attack scenario to 31,861% and 87,241%. In this attack scenario with a total of 1000 incoming connections, the web server is still able to serve requests even though it terminates incoming connections several times due to resource limitations.

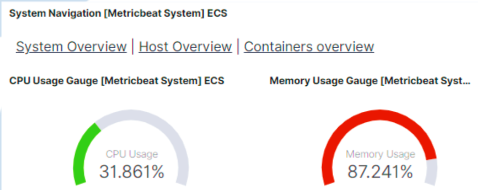


Fig. 10. CPU Usage and Memory Usage in Scenario 2

In testing scenario 3 as shown in Fig. 11, the web server is loaded with request traffic of 1000 connections originating from 2 PCs using Slowloris with a duration of 20 minutes. In total there are 2000 connections that make concurrent requests to the web server.

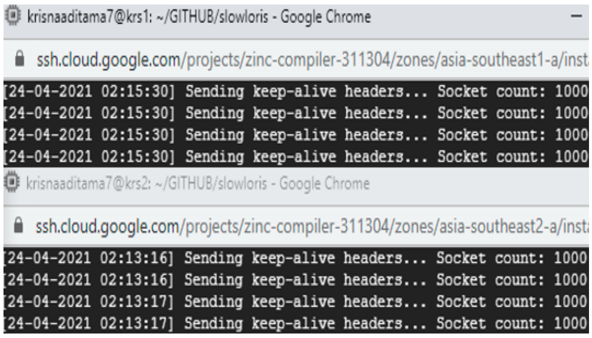


Fig. 11. Slowloris attack with 1000 connections from 2 PCs

Fig. 12. Web Server Log Results in Scenario 3

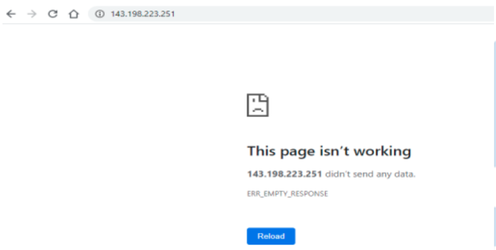


Fig. 13. Inaccessible Web page

From the observations shown in Fig. 12, it can be seen that the web server has difficulty handling the many incoming requests collectively so that the service does not run well, the log graph indicator is not recorded normally and the web page display is not available as shown in Fig. 13

In the last attack scenario, although the total attack was 2 times larger than the previous one, CPU Usage and Memory Usage did not increase significantly, at 29.462% and 88.445% as shown in Fig. 14, this is due to the web server service that is not running to handles the number of other incoming connections. So that there is no visible increase in the number of cpu and memory usage even though the system is experiencing problems, where web pages cannot be accessed.

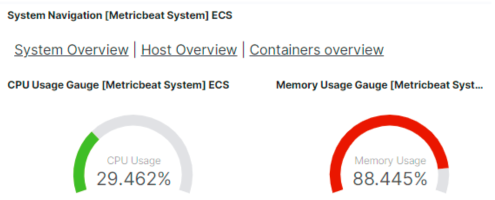


Fig. 14. CPU Usage and Memory Usage in Scenario 3

Based on Table IV, the web server test shows that the web service is still running and accessed normally when the number of incoming requests is 500 connections, when the number of requests given reaches 1000 connections most of the connection requests are still able to be served even though there are several times the request fails to be handled, due to limited resources. In testing with a total of 2000 requests, then the web server was unable to serve its services, resulting in the web page not functioning at all. This means that the Nginx web server resource by default is only able to handle requests for no more than 1000 simultaneous incoming connections, in the experimental conditions on the server specifications that are made as in Table I. In general Nginx as a web server is better able to cope with types of attacks such as Slowloris, but further configuration is needed to be able to serve more connections to the server.

TABLE IV. ATTACK IMPACT

Scenario	Total	Impact of Web Server
1	500 Connection	Service running normally
2	1000 Connection	Several times the service stopped
3	2000 Connection	Service stopped running

B. *Attack Origin Testing*

The second test aims to see the log data of the detected attack origin, in the form of the country of origin of the attack and attempts to access server logins from outside parties. The [Filebeat Nginx] section of the Elastic Common Schema (ECS) Overview, is capable of displaying a source map and details of the origin of the attack. From the resulting display as shown in Fig. 15, it was identified that the attack that was launched came from the Singapore area, this is in accordance with the experimental scenario where the attacker's PC used were from that region.

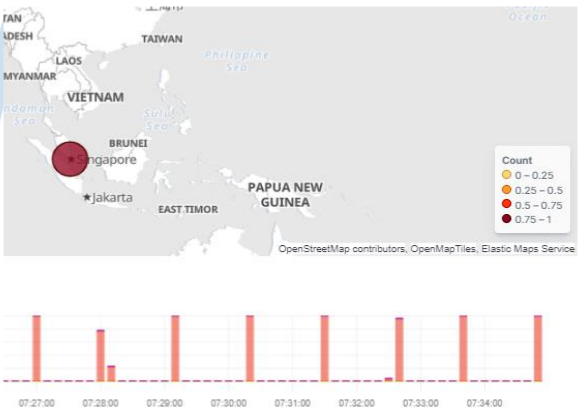


Fig. 15. Map of the origin of attack

Testing of login attempts to the server were carried out by entering a username and password at random to see if the Elastic Stack can record these attempt, statistical results can be seen in the [Filebeat System] section of ECS SSH login attempts. From the graphic shown in Fig. 16, there are a lot of attacks from outside parties trying to enter the system, whose IP addresses were detected from various countries. It can be seen that the largest number of attacks came from China territory. This proves earlier allegations that many of the attacks originated in that region.

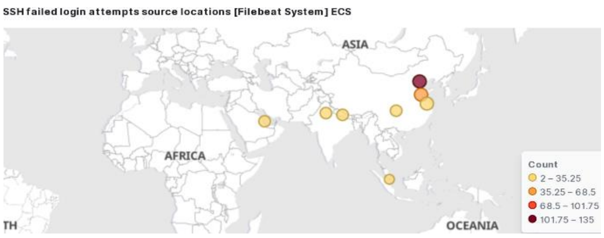


Fig. 16. Map of access for server logins

Several login attempts were recorded every time with various random user combinations, as the sample shown in Table V, these records are always updated on the [Filebeat System] SSH login attempts ECS.

TABLE V. SSH LOGIN ATTEMPTS

Time	user.name	source.ip
Apr 24, 2021@ 10:14:07.000	admin	161.97.183.201
Apr 24, 2021@ 10:14:10.000	camera	103.150.136.128
Apr 24, 2021@ 10:14:15.000	bertrand	42.51.9.162
Apr 24, 2021@ 10:14:16.000	user1	14.02.74.99
Apr 24, 2021@ 10:14:18.000	bertrand	42.51.9.162

IV. CONCLUSION

The results obtained are the process of measuring and analyzing server capabilities is made easier by using Elastic Stack in log management, with statistics and visuals generated. Based on the test results, a server with specifications of 2 CPUs and 4GB of memory is able to handle requests of 500 to 1000 connections, but the service will have difficulty until it stops running when the total incoming requests reaches 2000 connections. The experimental results show that the CPU Usage obtained when the attack was carried out with scenarios of 500 and 1000 connections was 26% and 31%, respectively. However, when the attack is carried out with a scenario of 2000 connections, the web server service is unable to serve requests and the CPU Usage is 29%. As for Memory Usage, every increase in connection requests results in an increase in memory consumption from 86% to 88%. Another result is that the source of the attack is identified according to the domicile of the attacker's simulated PC, namely Singapore. In addition, there is also a list of login access, both internal and external, using random users and passwords to enter the server system. The addition of the login authentication feature on the Elastic Stack and the use of Nginx as a web server is highly recommended to minimize the impact of the Slowloris attack.

REFERENCES

[1] C. Tarigan and D. Angela, "Sistem Pengawasan Kinerja Jaringan Server Web Apache dengan Log Management System ELK (Elasticsearch , Logstash , Kibana)," *J. Telemat. Ed. Ind. Eng. Semin. Call Pap.*, vol. 1, no. 1, pp. 7–14, 2018.

[2] A. F. Rochim, M. A. Aziz, and A. Fauzi, "Design Log Management System of Computer Network Devices Infrastructures Based on ELK Stack," *ICECOS 2019 - 3rd Int. Conf. Electr. Eng. Comput. Sci. Proceeding*, no. May 2020, pp. 338–342, 2019, doi: 10.1109/ICECOS47637.2019.8984494.

[3] D. Berman, "The Complete Guide to the ELK Stack," 2020. https://logz.io/learn/complete-guide-elk-stack/.

[4] J. Ellingwood and V. Kalsin, "How To Install Elasticsearch, Logstash, and Kibana (Elastic Stack) on Ubuntu 18.04," 2018. https://www.digitalocean.com/community/tutorials/how-to-install-elasticsearch-logstash-and-kibana-elastic-stack-on-ubuntu-18-04.

[5] K. Lotz, "Integrating the Elastic Stack into ExplorViz to Collect Logs and Runtime Metrics," 2019.

[6] T. Diotalevi *et al.*, "Collection and harmonization of system logs and prototypal Analytics services with the Elastic (ELK) suite at the INFN-CNAF computing centre," *Proc. Sci.*, vol. 351, pp. 0–14, 2019, doi: 10.22323/1.351.0027.

[7] S. Alghfeli, Z. Alhadrami, M. Alghfeli, N. Albloushi, and A. Alfaresi, "Bayyinah, A Log Analysis Forensics Tool," *Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019*, no. Ii, pp. 845–849, 2019, doi: 10.1109/AICAI.2019.8701405.

[8] M. Harikanth and P. Rajarajeswari, "Malicious event detection using ELK stack through cyber threat intelligence," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 7, pp. 882–886, 2019.

[9] W. Sholihah, S. Pripambudi, and A. Mardiyono, "Log Event Management Server Menggunakan Elastic Search Logstash Kibana (ELK Stack)," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 2, no. 1, pp. 12–20, 2020, doi: 10.35746/jtim.v2i1.79.

[10] M. Bajer, "Building an IoT data hub with elasticsearch, Logstash and Kibana," *Proc. - 2017 5th Int. Conf. Futur. Internet Things Cloud Work. W-FiCloud 2017*, vol. 2017-Janua, no. August 2017, pp. 63–68, 2017, doi: 10.1109/FiCloudW.2017.101.

[11] S. M. Helalat, "An Investigation of the Impact of the Slow HTTP DOS and DDOS attacks on the Cloud environment," Blekinge Institute of Technology, 2017.

[12] A. Y. Chandra, "Analisis Performansi Antara Apache & Nginx Web Server Dalam Menangani Client Request," *J. Sist. dan Inform.*, vol. 14, no. 1, pp. 48–56, 2019, doi: 10.30864/jsi.v14i1.248.

[13] M. Arman, "Metode Pertahanan Web Server Terhadap Distributed Slow HTTP DoS Attack," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 56–70, 2020, doi: 10.35957/jatisi.v7i1.284.

[14] Y. Li *et al.*, "A cloud-based framework for large-scale log mining through Apache Spark And Elasticsearch3," *Appl. Sci.*, vol. 9, no. 6, 2019, doi: 10.3390/app9061114.

[15] F. Abd. Hadi, M. Ahyar, and I. Syamsuddin, "ELK : Teknologi Mesin Pencari Big Data Terdistribusi," *Semin. Nas. Tek. Elektro dan Inform. 2018*, no. September, pp. 333–338, 2018.

[16] V. K. Et. al., "Twego Trending: Data Analytics Based Search Engine Using Elasticsearch," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 1S, pp. 246–251, 2021, doi: 10.17762/turcomat.v12i1s.1764.

[17] K. Yamnual, P. Phunchongharn, and T. Achalakul, "Failure detection through monitoring of the scientific distributed system," *Proc. 2017 IEEE Int. Conf. Appl. Syst. Innov. Appl. Syst. Innov. Mod. Technol. ICASI 2017*, pp. 568–571, 2017, doi: 10.1109/ICASI.2017.7988485.

Academic Information Systems and Recommendations using AHP at SMA Islamic Center Tangerang

Alvira Putri Yudini[1], Calandra Alencia Haryani[2], Andree E. Widjaja[3], Hery[4], Suryasari[5]
[1,2,3,4]Information Systems, Universitas Pelita Harapan, Tangerang, Indonesia
[5]Department of Informatics, Multimedia Nusantara University, Tangerang, Indonesia
[1]alvirap7@gmail.com, [2]calandra.haryani@uph.edu, [3]andree.widjaja@uph.edu, [4]hery.fik@uph.edu, [5]suryasari@umn.ac.id

Accepted 24 October 2021
Approved 4 January 2022

***Abstract*— Nowadays, technology is indispensable to solve problems. Technology can facilitate communication without thinking about distance, space, and time. Currently, the advancement of technology is useful for most people in carrying out their activities, especially in the field of education. Islamic Centre High School Tangerang is one of the schools that require information systems to manage data as it still uses the manual method.**

This research was specifically conducted to assist t Islamic Centre High School Tangerang in building a system for managing school data, so that it can provide recommendations to students to decide on a study program to a higher level of education, as well as to help parents monitoring their children activities. The methodology used in this study is qualitative methods, RAD methodology and the Prototyping method. Meanwhile, the method of recommending study programs for students used the Analytical Hierarchy Process method, that is by comparing the value of criteria consisting of accreditation, majors, grades, and comparison of alternative values consisting of study programs and universities. The system was tested using black box testing method. The result of this study is a web-based academic information systems with recommendation feature. The system can display a ranking of student study program recommendations based on a comparison of criteria and alternative values using the Analytical Hierarchy Process Method.

Index Terms— Recommendation, Academic Information System, Analytical Hierarchy Process (AHP), prototyping

I. Introduction

Technology can be used as a tool to solve a problem or as a communication tool [1]. The use of technology provides convenience, such as processing data, sharing information and doing business. Most of the technology has been widely used in the fields of business, entertainment, industry, communication, and education. When technology has not developed as it is today, all activities are done manually, in so doing activities become slow and thus require a lot of time. Education is one of the fields that require technology. Current technological advancements in education field are expected to assist educators in providing convenience in managing data [2][3]. Therefore, with the development of current technology, it is very helpful for most humans in carrying out their activities, especially in the field of education which requires an accurate information system to minimize errors [4], hence information systems are needed [5].

Information systems used in the field of education are called academic information systems. Academic information system is software used to present information and organize administration related to academic activities [6]. The services provided are as follows: management of student data, schedules, grades, and report cards. An educational institution is a place where the teaching and learning process takes place, including education in the family, school, and community. In the current era of globalization, educational institutions will be able to apply information technology to support academic activities such as data processing and presenting academic information quickly and accurately, but there are still some schools that use manual methods in presenting information [7][8].

This is mainly due to the lack of qualified teachers in the IT field. Based on the Central Statistics Agency which has conducted a survey in 4,014 schools spread across 34 provinces, there are results that the use and utilization of information and communication technology for the proportion of teachers who have qualifications in the IT field at the high school level and the equivalent is 14.43% [9]. One of the educational institutions that still uses manual methods in managing their data is Islamic Center High School Tangerang. Currently, they still use manual methods in carrying out academic activities such as taking attendance using paper, using grade books to store student grades, calculating them manually by asking each category weight to the respective subject teacher, calculating it and then entering it into the report card. Based on interviewed with headmaster, there are frequent errors in the calculation of grades and there are some students who complain because of the incorrect spelling of names on the report card because of manual tasking. So, an academic information system is purposefully created to assist SMA Islamic Center Tangerang in processing academic data properly.

The result can help SMA Islamic Center Tangerang to managing student and teacher data, viewing class schedules, viewing grades, downloading test cards, and printing student report cards. The required academic information system using Analytical Hierarchy Process (AHP) method can also help students in class X and XI to get recommendation for the study program they are interested in or want when they go to a higher level of education. AHP method used for this research because this method can give teacher privilege to manage category weight based on their preference for personalization recommendation study program result.

II. Method

The system development methodology used for the Academic Information System is the Rapid Application Development (RAD) using prototyping method. RAD is a development method a system in a relatively short time. Advantages obtained with using the RAD methodology is faster, more accurate, and more cost-effective inexpensive [10]. The Prototyping method is a system model that help development is done quickly and easily. The Prototyping method provides how to develop a more effective and efficient system through discussion, exploration, experimentation, and improvement over and over again [10]. The recommendation system for this webiste is using Analytical Hierarchy Process (AHP). The AHP was first developed and introduced in 1970 by Thomas L. Saaty, a Professor of Mathematics from the University of Pittsburgh [11]. AHP is one of the methods of Multi-Criteria Decision Making (MCDM). AHP is a measurement theory used to derive ratio scales from discrete and continuous pairwise comparisons [1] [11][12].

The AHP method is carried out by forming a problem structure and then compiling a hierarchical structure of the problem starting from the goal, then the criterion variable, then followed by alternative variables. The AHP method is widely used to prioritize options with many criteria, but its application has been widely used as an alternative model of cost benefits, forecasting and others [13]. In the field of education, the AHP method can help in making decisions about choosing a study program [14].

In solving problems with the AHP method, there are several basic principles, namely, decomposition (principles of compiling a hierarchy), comparative judgment, synthesis of priority, logical consistency [15][16][17]. The first step in determining the priority of elements in a decision problem is to make pairwise comparisons. The AHP method uses pairwise comparison between one factor and other factor to produce factor weights and between one alternative to another alternative for evaluation factor. The factor uses pairwise comparison matrix as with 1-9 scale [1]. The elements are compared in pairs against a specified criterion. Table 1 pairwise comparison presented in the form of a matrix.

Table I. Is The Scale Used to Fill In the Pairwise Comparison Matrix

Level Of Importance	Definition
1	Equality Important
3	Moderately More Important
5	Slightly More Important
7	Very Strongly More Important
9	Extremely More Important
2, 4, 6, 8	Neutral value between slightly more important and more important

The pairwise comparison matrix can be normalized with the following steps [18]:

1. Add up the values of each column of the pairwise comparison matrix: $\sum_{i=1}^{n} a_{ij}$ for $i, j = 1,2, \dots, n$
2. Divide the value a_{ij} in each column by the sum of the values in the column: $a_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}$ for $i, j = 1,2, \dots, n$.
3. Add up all the values of each row of the normalized matrix and divide by the elements of each row. The result of the division shows the priority value for each element.

Inconsistency often occurs in pairwise comparison assessments. The consistency of the paired assessment was evaluated by calculating the Consistency Ratio (CR). When determining if CR = 0.1 then the results of the assessment are said to be consistent. The formulation to calculate the Consistency Ratio is as follows:

$$CR = \frac{CI}{RI} \quad (1)$$

Description:
CI = Consistency Index

RI = Random Consistency Index

The formula for calculating CI is as follows:

$$CI = \frac{(\lambda_{max} - n)}{n-1} \qquad (2)$$

Description:
λ_{max} = the maximum value of the eigenvalues of order n.

The maximum eigen value is obtained by adding up the result of the comparison matrix multiplication with the main eigenvector (priority vector) and dividing it by the number of elements. The CI value will be meaningless if there is no reference to state whether the CI shows a consistent or inconsistent matrix.

Ordo Matriks	1,2	3	4	5	6	7	8	9	10	11	12	13
RI	0	0,52	0,89	1,11	1,25	1,35	1,4	1,45	1,49	1,51	1,54	1,56

Fig. 1.Random Index (RI) value obtained by Saaty.

Any scalar multiple kv of the eigenvector v that includes $\lambda$is also an eigenvector because:

$$A(kv) = k(Av) = k(\lambda v) = \lambda(kv) \qquad (3)$$

To achieve the eigenvalues of the matrix A of size $n \times n$. then it can be written in the following equation:

$$Av = \lambda v \qquad (4)$$

Or equivalently:

$$(\lambda I - A)v = 0 \qquad (5)$$

For λ to be an eigenvalue, there must be a non-zero solution of the above equation. However, the above equation will have a non-zero solution if and only if:

$$det(\lambda I - A)v = 0 \qquad (6)$$

The above equation is called the characteristic equation A, the scalar that satisfies the above equation is the eigenvalue of A. It is known that the value of the ratio of element A_i to element A_j is a_{ij}, so theoretically the matrix has the opposite positive characteristic, namely $a_{ij} = \frac{1}{a_{ij}}$. The weight sought is expressed in vector $w = (w_1, w_2, w_3, \dots, w_n)$. The value of w_n represents the weight of criterion A_n against the entire set of criteria in the sub-system. If a_{ij} represents the degree of importance of factor i to factor j and a_{ik} represents the degree of importance of factor j to factor k then for the decision to be consistent, the importance of factor i to factor k must be equal to $a_{ij} \cdot a_{jk}$or if $a_{ij} \cdot a_{jk} = a_{ik}$ for all i, j ,k. For a matrix consistent with vector w, then the element aij can be written:

$$a_{ij} = \frac{w_i}{w_j}; \forall i,j = 1,2,3,\dots,n \qquad (7)$$

So, the consistency matrix is as follows:

$$a_{ij} \cdot a_{jk} = \frac{w_i}{w_j} \cdot \frac{w_j}{w_k} = \frac{w_i}{w_k} = a_{ik} \qquad (8)$$

As described above, the pairwise comparison matrix can be broken down into:

$$a_{ij} = \frac{w_i}{w_j} = \frac{1}{w_j/w_i} = \frac{1}{a_{ji}} \qquad (9)$$

From the above equation, we get:

$$a_{ij} \cdot \frac{w_j}{w_i} = 1 \qquad (10)$$

From the above equation, it can be seen that the consistent pairwise comparison matrix is:

$$\sum_{i.j=1}^{n} a_{ij} \cdot a_{ij} \cdot \frac{1}{w_{ij}} = n; \forall i,j = 1,2,3,\dots,n \quad (11)$$

$$\sum_{i.j=1}^{n} a_{ij} \cdot a_{ij} = nw_{ij}; \forall i,j = 1,2,3,\dots,n \qquad (12)$$

The above equation is equivalent to the form of a matrix equation:

$$A \cdot w = n \cdot w \qquad (13)$$

In matrix theory, the above formulation explains that w is the eigenvector of matrix A with the Eigen values of n. Note that n is the dimension of the matrix itself. In the form of a matrix equation, it can be written as follows:

$$\begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} \cdots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} \cdots & \frac{w_2}{w_n} \\ \vdots & \vdots \ \ddots & \vdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} \cdots & \frac{w_n}{w_n} \end{bmatrix} \qquad (14)$$

But in reality, it cannot be guaranteed that:

$$a_{ij} = \frac{a_{ik}}{a_{jk}} \qquad (15)$$

One of the causes of this happening,, is the human element (decision maker) because it is not always consistent and absolute in expressing preferences for the elements being compared. In other words, the assessment given to each element of the problem at a hierarchical level may be inconsistent.

III. Result and Discussion

The following are the steps in calculating AHP:

A. Create a recommendation hierarchy

The first step in calculating AHP is to create a hierarchy consisting of three levels, the first level of objectives, the second level of assessment criteria and the third level of alternative choices and assessments. The factor that determines the results of the recommendations is by determining the weights on the criteria and alternatives. The weight of the assessment is carried out first to create a pairwise comparison matrix that describes the effect of each criterion on the alternatives. Figure 2 is a hierarchy of recommendations for the "Information Academic Islamic Center High School" system.

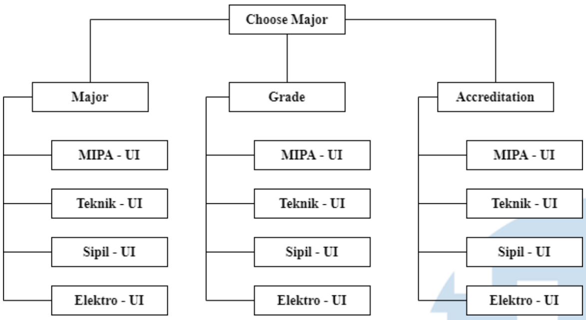


Fig. 2. Recommendation Hierarchy

The data in Figure 2 is the data entered by the user. Users can add their own criteria and alternatives. The criteria in Figure 2 were added based on the Ministry of Education and Culture's assessment for SNMPTN.

B. Do pairwise comparisons

After creating the hierarchy, the next step is to perform pairwise comparisons based on the Saaty scale to get the weight of the criteria. Table 2 is a pairwise comparison to get the weight of the criteria.

TABLE II. COMPARISON TABLE OF PAIRED CRITERIA

	Grade	Accreditation	Major
MAJORS	1	1/2	3
ACCREDITATION	2	1	4
GRADE	1/3	1/4	1

C. Performing the calculation of the weight of the criteria

After performing pairwise comparisons, the next step is to calculate the weight of the criteria (priority vector). Figure 3 is the calculation of the weight of the criteria. The criterion matrix is obtained after doing a weight comparison, after that normalization of each matrix, according to the formula contained in the previously described equation. The results of the normalization of the matrix will be added every row, if you get the results, then the columns will be summed. The result of the sum will be calculated with the result of the column sum to get the average matrix.

$$\text{Criteria} = \begin{pmatrix} 1 & 0{,}5 & 3 \\ 2 & 1 & 4 \\ 0{,}33 & 0{,}25 & 1 \end{pmatrix} \text{Matrix Normalization Result} \rightarrow \begin{pmatrix} 0{,}3 & 0{,}29 & 0{,}38 \\ 0{,}6 & 0{,}57 & 0{,}5 \\ 0{,}1 & 0{,}14 & 0{,}13 \end{pmatrix} \text{Matrix Average Result} \rightarrow \ \mathbf{x} \begin{pmatrix} 0{,}3 \\ 0{,}6 \\ 0{,}1 \end{pmatrix}$$

$$\text{Total Column} = 3{,}33 \quad 0{,}25 \quad 8 \qquad\qquad 1 \quad 1 \quad 1$$

Fig. 3 Calculation of Criteria Weight

D. Check Consistency Ratio (CR)

The next step is to check the Consistency Ratio (CR) of the criterion paired comparison matrix. If CR > 0.1, the pairwise comparison must be repeated until CR <= 0.1. Before calculating the CR value, it must be calculated to get the Consistency Index (CI) value. The CI formula can be found in the previously described equation. Figure 4 is a calculation to get the CI value.

$$\overset{K}{\begin{bmatrix} 1 & 0{,}5 & 3 \\ 2 & 1 & 4 \\ 0{,}33 & 0{,}25 & 1{,}0 \end{bmatrix}} \overset{x}{\begin{bmatrix} 0{,}30 \\ 0{,}60 \\ 0{,}10 \end{bmatrix}} = \overset{Kx}{\begin{bmatrix} 0{,}90 \\ 1{,}60 \\ 0{,}35 \end{bmatrix}} = \lambda_{max} \overset{x}{\begin{bmatrix} 0{,}30 \\ 0{,}60 \\ 0{,}10 \end{bmatrix}}$$

$$\lambda_{max} = average\left\{\begin{pmatrix} 0{,}90 \\ 0{,}30 \end{pmatrix}, \begin{pmatrix} 1{,}60 \\ 0{,}60 \end{pmatrix}, \begin{pmatrix} 0{,}35 \\ 0{,}10 \end{pmatrix}\right\} = 3{,}06$$

$$CI = \frac{(\lambda_{max} - n)}{n-1} = \frac{(3{,}06 - 3)}{3-1} = 0{,}03$$

Fig. 4. Calculation of CI

After getting the CI value contained in Figure 4, you can directly perform calculations to get the CR value. Figure 5 is a calculation to get the results of the CR value.

$$CR = \frac{CI}{IR} = \frac{0{,}03}{0{,}58} = 0{,}05$$

Fig. 5. Calculation of CR

Based on the calculation results obtained in Figure 5 the results of CR = 0.05, which means CR <= 0.1, so it is consistent. The IR value is obtained based on Figure 1 which corresponds to the order of the matrix. Figure 6 is a new hierarchical arrangement added with the weights of each criterion.

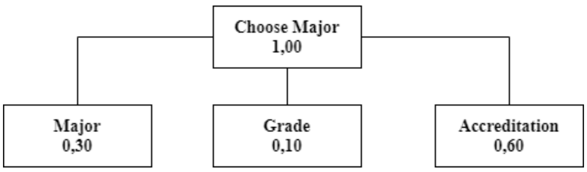


Fig. 6. Hierarchy Added with Weighted Criteria Calculation

E. Calculation of alternative weights

The next step is to perform calculations for alternative weights. Each alternative performs calculations based on existing criteria. Based on the hierarchy in Figure 7, the weight calculations that must be carried out are alternatives for majors' criteria, alternatives for value criteria, and alternatives for accreditation criteria. Figure 8 is the result of calculating alternative weights based on criteria.

Major	MIPA-UI	Teknik-UI	Sipil-UI	Elektro-UI	Priority Vector
MIPA-UI	1	1/4	4	1/6	0,13
Teknik-UI	4	4	1	1/4	0,24
Sipil-UI	1/4	1/4	1	1/5	0,07
Elektro-UI	6	4	5	1	0,56

Accreditation	MIPA-UI	Teknik-UI	Sipil-UI	Elektro-UI	Priority Vector
MIPA-UI	2	2	5	1	0,38
Teknik-UI	1	1	3	2	0,29
Sipil-UI	1/3	1/3	1	1/4	0,07
Elektro-UI	1	1/2	4	1	0,26

Grade	MIPA-UI	Teknik-UI	Sipil-UI	Elektro-UI	Priority Vector
MIPA-UI	1	2	5	1	0,38
Teknik-UI	1/2	1	3	2	0,29
Sipil-UI	1/5	1/3	1	1/4	0,07
Elektro-UI	1	1/2	4	1	0,26

Fig. 7. Alternative Calculations

Figure 9 is a new hierarchical arrangement added with criterion weights and alternative weights whose values are obtained from the results of calculations in Figure 8.

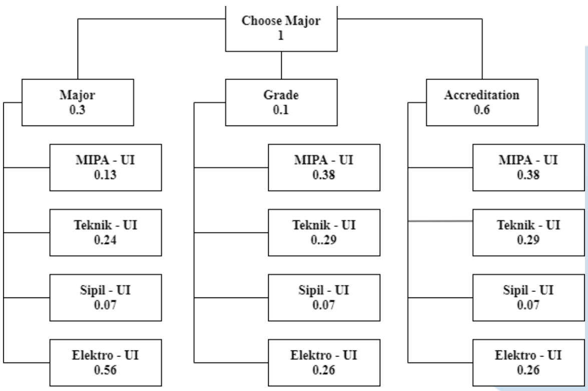


Fig. 8. Hierarchy with Calculated Weights

F. Calculating alternative ranking

The final step is to calculate the ranking of alternatives. The alternative ranking results are obtained by adding up the multiplication of each alternative weight with the weight of the corresponding criteria. Figure 9 is a calculation to get the ranking results.

$$\begin{array}{l} \\ \text{MIPA-UI} \\ \text{Teknik-UI} \\ \text{Sipil-UI} \\ \text{Elektro-UI} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Major} & \text{Accreditation} & \text{Grade} \end{array} \\ \begin{pmatrix} 0{,}13 & 0{,}38 & 0{,}38 \\ 0{,}24 & 0{,}29 & 0{,}29 \\ 0{,}07 & 0{,}07 & 0{,}07 \\ 0{,}56 & 0{,}26 & 0{,}26 \end{pmatrix} \end{array} \times \begin{pmatrix} 0{,}3 \\ 0{,}6 \\ 0{,}1 \end{pmatrix} = \begin{pmatrix} 0{,}8 \\ 0{,}5 \\ 0{,}06 \\ 0{,}4 \end{pmatrix}$$

Fig. 9. Alternative Rank Calculations

From Figure 9 it can be concluded that from the calculation of the criteria weights and the alternative weights the ranking results are as follows:

1. MIPA – UI with a result of 0,80.
2. Teknik – UI with a result of 0,50.
3. Sipil – UI with a result of 0,06.
4. Elektro – UI with a result of 0,4

This Academic Information System and Recommendations website is build using Indonesian Language can be accessed by 5 actors: admin, teacher, homeroom teacher, student, and parent. This website provides menus for manage report cards, manage self-development scores, manage grades, manage attendance, access recommendations, manage recommendations, change passwords, manage exam schedules, download test cards, manage user data, manage grade criteria, register users, manage schedules, manage subjects, manage the school year, manage classes, manage school payment information, view schedules, view report cards, access student grades, and access attendance. Each actor has a different menu access in this website. Figure 10 until Figure 15 shows a snapshot of some of the academic information system interfaces and recommendations that have been made. Figure 10 shows the system view for dashboard.

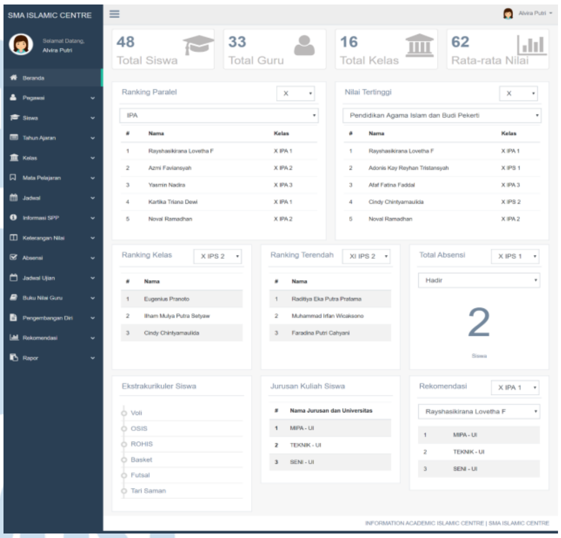


Fig. 10. Dashboard Page

Figure 11 shows the system view for comparison criteria value for recommendations.

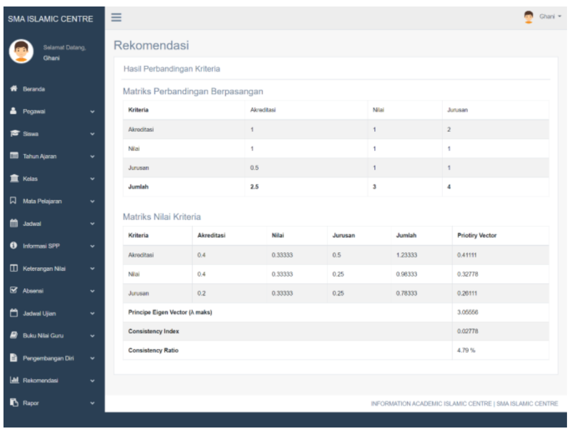


Fig. 11. Comparison Criteria Recommendations Page

Figure 12 shows the system view forrecommendation.

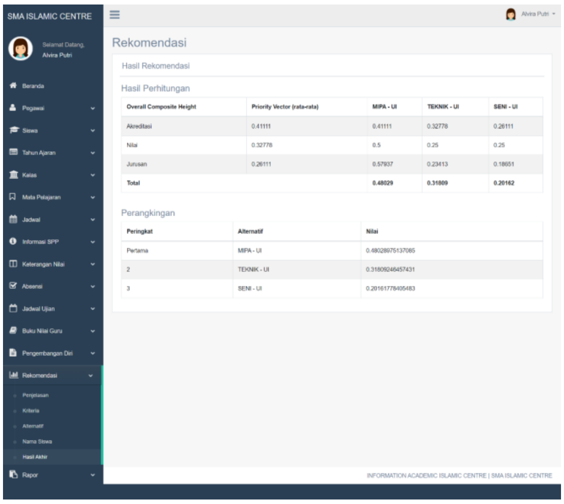


Fig. 12. Recommendations Page

Figure 13 shows the system view for report.

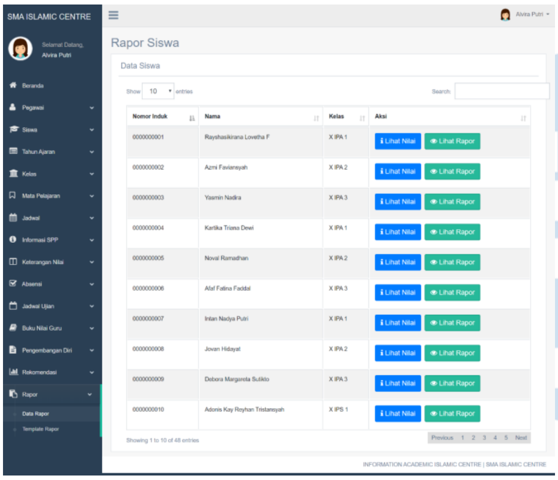


Fig. 13. Report Page

Figure 14 shows the system view for self-development scores.

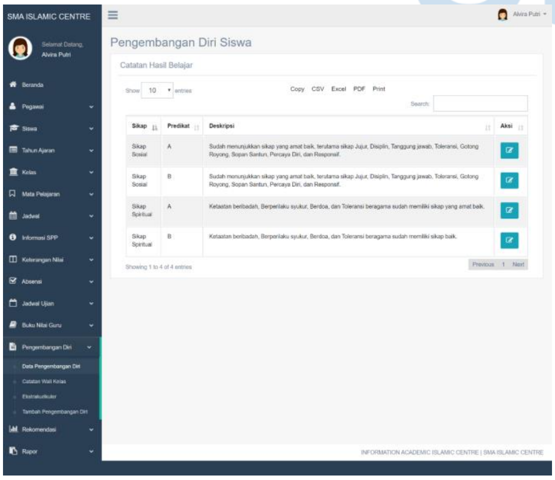


Fig. 14. Self-Development Scores Page

Figure 15 shows the system view for scheduling

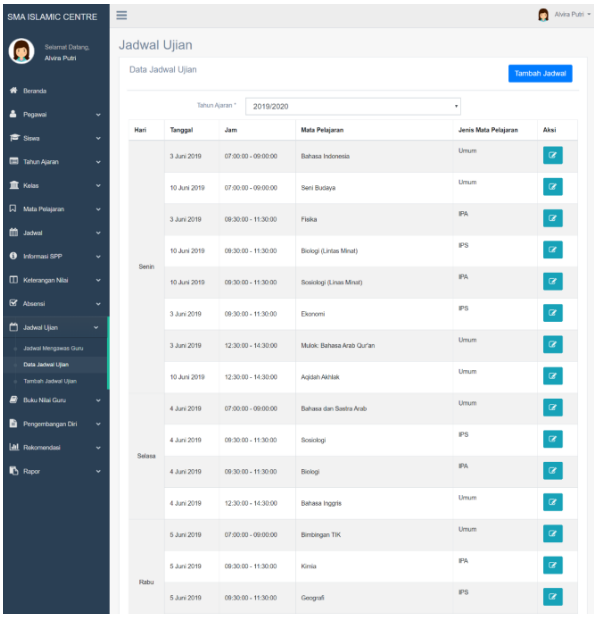


Fig. 15. Scheduling Page

This website uses Blackbox testing method for system functional testing. This website has successfully passed the testing phase on each menu. Table 3 shows an example of Blackbox testing for the criteria comparison page for major recommendations.

TABLE III. BLACKBOX TESTING CRITERIA COMPARISON PAGE

STEP	Test Action	Expected Results	Status
1	Admin chooses to create comparison	System displays the comparison form	Pass
2	Admin do comparison with determine past weight, then click result	System displays result of comparison and save it to database	Pass

IV. CONCLUSION

The result of this study is an academic information system for SMA Islamic Center. The conclusions obtained are based on the result of system and Blackbox testing for the user:

- The "Information Academic SMA Islamic Center" system has been successfully designed and developed using the website which can be accessed by homeroom teachers, teachers, homeroom students, students, and admins with different views according to their respective roles.
- The system can also display the grades of each student in the middle of the semester, so that each student can improve their learning for grade

promotion. The system for viewing report card data can also be printed in PDF form, making it easier for students and homeroom teachers when distributing report cards or when they need a hardcopy of report cards.

- The student data view system can also help each student if there is an incorrect personal data, but not all data can be changed by students, only some data can be changed by themselves. The system also helps students to view the student's class schedule, and the system can also print the student's class schedule.

- Home on the system can help parents/guardians of students in monitoring their children in the field of education. Parents/guardians also get access to see each student's grades.

- Recommendations for majors can be combined with other recommendation methods such as linear regression

REFERENCES

[1] R. Desanti, C. Supit, and A. Widjaja, "Aplikasi Perekrutan dan Penilaian Karyawan Berbasis Web pada PT. XYZ", Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol. 8, no. 2, pp. 74-80, Dec. 2017.

[2] X. Wang, "Design and Implementation of College English Teaching Quality Evaluation System Based on Analytic Hierarchy Process," 2020 International Conference on Computers, Information Processing and Advanced Education (CIPAE), 2020, pp. 213-216, doi: 10.1109/CIPAE51077.2020.00062.

[3] F. Fadhilah, "Pengembangan Sistem Informasi Akademik berbasis Web di SMA Negeri 1 Banjarharjo," Yogyakarta, 2016

[4] T. R. Sahroni and H. Ariff, "Design of analytical hierarchy process (AHP) for teaching and learning," 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), 2016, pp. 1-4, doi: 10.1109/KICSS.2016.7951412.

[5] M. Susanti, "Perancangan Sistem Informasi Akademik Berbasis Web pada SMK Pasar Minggu Jakarta," Jurnal Informatika BSI, vol. III, no. 1, pp. 91-99, 2016.

[6] Y. Sari, S. D. Ariyanto and S. B. Utomo, "Analysis and Design of Reengineering Academic Information System Astra Polytechnic of Manufacture in Interconnection Network," 2018 International Conference on Applied Science and Technology (iCAST), 2018, pp. 577-583, doi: 10.1109/iCAST1.2018.8751520.

[7] Nurbojatmiko, A. Fajar Firmansyah, Q. Aini, A. Saehudin and S. Amsariah, "Information Security Awareness of Students on Academic Information System Using Kruger Approach," 2020 8th International Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1-7, doi: 10.1109/CITSM50537.2020.9268795.

[8] F. Suryandani, B. dan D. Maryono, "Pengembangan Sistem Informasi Akademik Berbasis Web sebagai Sistem Pengolahan Nilai Siswa di SMK Negeri 1 Kudus," Jurnal Ilmiah Pendidikan Teknik Kejuruan (JIPTEK), vol. X, no. 1, pp. 71-82, 2017.

[9] T. Sutarsih dan A. N. Hasyyati, "Statistik Penggunaan dan Pemanfaatan Teknologi Informasi dan Komunikasi", Jakarta: BPS-Statistics Indonesia, 2018.

[10] A. Dennis, B. H. Wixom dan D. Tegarden, System Anlysis & Design; an Object Oriented Approach with UML fifth edition, Hoboken: John Wiley & Sons. Inc, 2015.

[11] T. I. Saaty, "Pengambilan Keputusan bagi ParaPemimpin: Hierar-ki Analitik untuk Pengambilan Keputusan dalam Situasi yang Kompleks", Jakarta: Pustaka Binaman Pressindo, 1991.

[12] M. A. Kossukhina, "Modeling of the innovation infrastructure of the university, based on the analytic hierarchy process," 2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM), 2016, pp. 509-511, doi: 10.1109/SCM.2016.7519829.

[13] S. Latifah, "Prinsip-Prinsip Dasar Analytical Hierarchy Process," *Studi Kasus Fakultas Pertanian Universitas Sumatera Utara (UsU)*, pp. 1-5, 2005.

[14] A. M. Syafar, "Sistem Pengambilan Keputusan Memilih Program Studi di UINAlauddin berbasis Web dengan Metode Analytic Hierarchy Process (AHP)," JurnalINSTEK (Informatika Sains dan Teknologi), vol. III, no. 2, pp. 309-318, 2018.

[15] M. Sri, "Riset Operasi", Edisi. Kedua, Jakarta: Lembaga Penerbit Fakultas Ekonomi Universitas Indonesia, 2004.

[16] M. Rahmayu dan R. K. Serli, "Sistem Pendukung Keputusan Pemilihan Jurusanpada SMK Putra Nusantara Jakarta menggunakan Metode Analytical HierarchyProcess (AHP)," Jurnal SIMETRIS, vol. IX, no. 1, pp. 551-564, 2018.

[17] D. Khuntari dan R. Ferdiana, "Sistem Rekomendasi Pemilihan Jurusan PerguruanTinggi dengan Pendekatan User Preference dan Analytic Hierarchy Process,"Seminar Nasional Teknologi Informasi dan Multimedia STMIK AMIKOMYogyakarta, pp. 1-6, 2015

[18] S. Latifah, "Prinsip-Prinsip Dasar Analytical Hierarchy Process," Studi KasusFakultas Pertanian Universitas Sumatera Utara (UsU), pp. 1-5, 2005

ISSN 2085-4552

Test Case Analysis with Keyword-Driven Testing Approach Using Katalon Studio Tools

Case Study: Angkasa Website

Reynaldi Prama Octavially[1], Rosa Reska Riskiana[2], Kusuma Ayu Laksitowening[3], Dana Sulistyo Kusumo[4], Monterico Adrian[5], Nungki Selviandro[6]
School of Computing, Telkom University, Bandung, Indonesia
[1]reyoctavially@student.telkomuniversity.ac.id, [2]rosareskaa@telkomuniversity.ac.id, [3]ayu@telkomuniversity.ac.id, [4]danakusumo@telkomuniversity.ac.id, [5]monterico@telkomuniversity.ac.id, [6]nselviandro@telkomuniversity.ac.id

Accepted 15 December 2021
Approved 22 January 2022

***Abstract*— Testing a software is an important stage of a series of software development. Functional testing of each feature on the Angkasa Website is intended to try out the function to match the required specifications. To achieve a functional test result, there are elements of features on the web page that require keywords. These keywords are used to perform actions or actions in running a web page, these keywords will help in making Test Cases for the testing process. Because it takes the right keywords to test on the web. To overcome this problem, this study analyzes the use of the Keyword Driven Testing approach for making Test Cases through the Katalon Studio tools. Keyword Driven Testing is one of the concepts in ISO/IEC/IEEE 29119, namely Keyword Driven Testing in The Test Design Process. The results of the analysis show that making Test Cases with Keyword Driven testing is easier to understand and is fully supported by the Katalon Studio tools. However, when creating test cases, not all keywords can be added automatically, so they need to be added manually. The automated testing successfully applied 10 keywords to test 26 features on the Angkasa Website with PASSED results for all of its features, but in the manual testing there is 1 bug found on Personal Information page. The difference result is because of Katalon Studio only checks whether the photo appears or not, but doesn't check whether the photo is corrupted or not.**

***Index Terms*— Automated Testing; Katalon Studio; Keyword Driven Testing; Test Cases; Website.**

I. INTRODUCTION

Software testing is one of the most important stages in the manufacture and development of a software, this test is carried out to determine the quality and feasibility of the software being developed. Ensure that the software being developed meets the requirements and is free from bugs and errors [1]. Software Testing is very important because it serves to ensure the quality of a software [2].

Software Testing is the last stage that must be passed in software development, Software quality can only be known through the testing phase, through technological advances throughout the world, there have been many improvements in terms of techniques and methods used to test a software [3]. An important aspect of the testing phase is to make sure the output is correct or not [4].

Website is a software that is widely used for organizational needs as a place to find and share information, a website has become a very powerful communication tool and its success depends on its accessibility, SEO (Search Engine Optimization) and its usability [5]. But not a few websites have problems with some of their features so that the website can't be used fully.

Based on this explanation, all software that is being developed must pass the testing stage before being released and used by users so that the software can run properly and not have problems that can confuse and disappoint users.

There are many approaches that can be used in automated software testing, either using a programming language or using a tool that can simplify the testing process. One approach that can be used is Keyword-Driven Testing in The Test Process which is a concept from ISO/IEC/IEEE 29119. This concept aims to support the conversion of keyword-based test cases into various types of test scripts that can be automatically executed by tools used [6].

There are many methods that can be used in automated testing of software, one of which is using Behavior Driven Development (BDD), which is a method that describes what activities need to be carried out in software testing and has the advantage of displaying information so that it can be easily understood by members. Project team and business team [7]. But if there are changes to the website that is being tested, it will take a long time to change back the tests that have been done previously, in contrast to testing using Keyword Driven Testing. In Keyword-Driven Testing, all the keywords used for testing a website can be easily added, removed and modified so that if there are changes to the website, the test does not need to be done from scratch [8]. By using Keyword-Driven Testing, the tests created can be easier to maintain and allow for easy understanding so that all team members can work together to carry out the test [9]. Therefore, this paper aims to analyze the implementation of Keyword-Driven Testing to make test cases for testing the Angkasa Website, because the Angkasa website has many features that must go through a testing phase to ensure whether these features are in accordance with the requirements or not, for example, such as the payment feature which involves money, because if this feature is not tested it will harm many parties.

In addition, similar research explains that Keyword Driven Testing saves a lot of manual testing by automating the website testing process for various test cases and also every keyword that has been used can be reused to test different websites [8]. In research on the evolution of Keyword Driven Testing shows that Keyword Driven Testing has a complex design with several levels of abstraction and a design that supports reuse, more than 60% of keywords used have the potential to reduce changes required during evolution by up to 70% [9].

In this paper, the researcher will use Katalon Studio to test the Angkasa Website because it is very suitable for beginners who are not used to testing software, Katalon Studio is also easier to use than other tools because in Katalon Studio the testers do not need to write scripts manually, everything is automatic so testing will not take a long time. Katalon Studio provides various features and has detailed, easy-to-read reports that can be presented in various formats such as HTML, CSV and PDF and can be shared with team members at any time [1]. Katalon studio has features that make it easy to use Keyword-Driven Testing, namely the Record and Playback features. The Record and Playback feature is very easy to use for testing a software and also speeds up processing time, so this feature was chosen for testing on the Angkasa Website. The purpose of this research is to find out what keywords will be used for the Keyword Driven Testing method which will then be tested for all the features on the website, whether all these features are ready to use or need to be repaired, all of this can be done through automatic testing using Keyword Driven Testing with keywords that have been obtained using Katalon Studio Tools.

II. Related Work

A. Literature Review

On research [7]-[9] describes the implementation of BDD (Behavior Driven Development) which is integrated with automation testing tools in software development using Katalon Studio Tools. The method used is Keyword Driven Testing which uses keywords to perform a software test, because Keyword Driven Testing is very easy to add and update and does not require much time in the testing process.

Automated testing of software that requires less time and less cost than manual testing is a testing technique using a code to perform machine-driven testing [10]. In Automation Testing, all software testing activities are carried out automatically starting from development, verification, and testing using Automated Testing Tools [11]. Katalon studio is one of the tools that is often used to implement the Keyword Driven Testing method, because Katalon Studio has a feature that makes it easy to find keywords to use [1].

Based on the literature review that has been explained, this research will analyze the result of testing implementation using Keyword Driven Testing to make test cases which is part of BDD using Katalon Studio on Angkasa Website.

B. Keyword Driven Testing

Keyword Driven Testing has undergone a change which was originally in the form of Scripted Test Automation into a test table with various keywords in it which serves to perform testing on a software.

Keyword Driven Testing is an increasingly popular testing method or approach because it involves the creation of reusable, modular test components. In Keyword Driven Testing, the system testing functionality is packaged in a table and it contains step-by-step instructions for each test [6].

ISO (the International Organization for standardization) and IEC (the International Electrotechnical Commission) have established a special system for standardization in the world, in addition the bodies that are members of ISO or IEC also participate in the development of international standards. [12].

ISO/IEC/IEEE 29119 aims to define a set of internationally agreed standards for software testing that can be used by any organization performing software testing [12]. ISO/IEC/IEEE 29119-5 will explain the main concepts of Keyword Driven Testing.

In Keyword Driven Testing, testing can be carried out more easily and can also be easily corrected if there are changes to the software being tested, Keyword Driven Testing can also allow experts from different fields and backgrounds to work together in different places [9]. To automate testing, each keyword needs to be implemented in the software [13].

- Keyword Rank

  In Keyword Driven Testing, the keywords used consist of 2 levels, described as follows [13]:

  1. Low Level, at this level the keys are associated with a collection of one or more actions that explain what steps must be

taken in the testing process. Example: click, set text, select, and others.

2. High Level, at this level keywords require a set of input parameters, which are also included in the structure. Keywords and parameters form a high-level description of the actions associated with the test case. Example: Open browser, login, and others.

- Layers in Keyword Driven Testing

There are 3 layers in Keyword Driven Testing, explained as follows [13]:

1. Domain Layers

The keywords in Domain Layers correspond to a domain-related business or activity and reflect the terminology used by domain experts. Keywords developed at Domain Layers are generally not implementation dependent.

2. Test Interface Layers

The keywords in Test Interface Layers refer to the type of interface used in a particular test. The actions required to address the test items can usually be easily identified.

3. Multiple Layers

Multiple Layers serves as a container that combines several existing Layers namely Domain and Interface Layers, Multiple layers can help manage hierarchical keywords.

- Keyword Type

There are 2 types of keywords in Keyword Driven Testing. Described as follows:

1. Simple Keywords

The simple keyword is a keyword that is often used in Test Interface Layers, such as MenuSelect and PressButton.

2. Compound Keywords

Combined keywords are a set of keywords or a keyword that is composed of other keywords, combined keywords can be arranged in different layers.

- Keyword Identification

Keyword Driven Testing requires the identification and definition of keywords. There are several sources that can be used to identify and define keywords, including the following [13]:

1. Exploration Test

In exploratory testing, the examiner observes which steps are performed. The new keyword is defined by providing a meaningful name, if the sequence of steps can be used with different data, the keyword will take the appropriate parameters for that data [13].

2. Business Expert

Keywords can be determined by conducting interviews with business experts, this question can be “What should be done to verify the application?” or “what needs to be tested?”. The answers given by the experts will be used to identify keywords by finding terms that may appear frequently [13].

3. Interface Test

Keywords can be specified through the test interface, but because interface elements are limited and usually small, a number of keywords can be defined for these interface elements. This approach will define low-level keywords in the Test Interface Layers [13].

4. Documented test procedures and test cases

The available test procedures and test cases can also be used as material for determining keywords. If two or more of the keywords found refer to the same activity, those keywords will only be replaced with the one keyword that best describes the activity [13]. Here are the conditions for keywords [13]:

a. Keywords must contain a verb.

b. Keywords must use the imperative form.

c. Keywords should provide a description of a series of related actions.

d. The keyword description should be clear.

e. Keywords must be defined in such a way that they can be understood by stakeholders who will use them when designing tests.

f. Each keyword must be unique in its meaning within a framework.

- Implementation of Keyword Driven Testing

This section discusses several concepts that contribute to the successful implementation of Keyword Driven Testing, there are six concepts discussed in this section, namely [13]:

1. Identifying keywords.

2. Composing test cases

3. Keywords and data driven testing.
4. Modularity and refactoring.
5. Keyword driven testing in the test design.
6. Converting non-keyword driven cases into keyword driven testing.

C. *Katalon Studio*

Katalon studio is an automation testing tool developed by Katalon LLC released in January 2015 launched on Microsoft Windows, macOS, and Linux. [10]. Katalon studio is built on the Selenium Framework's open-source test automation, Appium using an IDE interface, built specifically for testing web-based, API, mobile, and desktop applications. [7].

Katalon Studio has a comprehensive set of features to implement [14], supports testing using keywords or Keyword Driven Testing which allows users who do not have experience in application testing to do it easily.

III. Methodology

This study applies the concept of Keyword Driven Testing, namely Keyword-Driven Testing in The Test Design as a reference for conducting automated testing on the Angkasa Website. The steps taken are data collection, automatic testing, analysis, and conclusions, as shown in Figure 1.

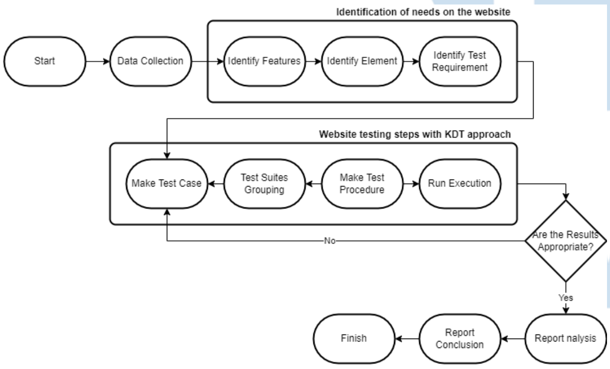


Fig. 1. Research Methods

A. *Test Design and Implement Process*

The following are some parts of the test design and implementation process used in the research shown in Figure 1.

- Identify Feature

  Identify feature on the Angkasa Website which will later be tested based on previously created documents. In this step, the examiner will note what features are on the Angkasa Website and will pass the testing stage.

- Identify Element

  A collection of special restrictions that contain functionality such as transaction features, functions, or structural elements that exist on the Angkasa Website. In this step, the examiner will record what elements are on the Angkasa Website to be processed in the testing phase.

- Identify Test Requirement

  Describe the test requirements at least once. What actually qualifies to be a covered requirement depends on each team's interpretation. In this step, the examiner will record what are the conditions for the test to be said to be successful or not.

Based on the three steps above, automatic testing will then be carried out using the Katalon Studio tools, namely the creation of a Test Case with a Keyword Driven Testing approach, derive Test Suite, test procedures, and program execution. The following are the steps carried out in the research.

- Make Test Case

  Test Cases are generated automatically either from specifications or code [15]. The test case that will be made focuses on positive cases that are adjusted to the acceptance criteria documents from the Angkasa Website. The Test Case is made by determining the keywords to be used in accordance with the Keyword Driven Testing approach, these keywords are determined based on the user story listed in the Software Requirements Specification. The following is an example of keywords obtained from 3 features on the Angkasa Website, the keywords can be seen in table I.

TABLE I. Identification of Keywords

Feature	Expected Behavior	Keywords
Login	Can press the login button to enter the login page.	Click
	Can fill in the username.	Set text
	Can fill in the password.	Set encrypted
	Can login by pressing the login button.	Click
Payments	Can choose the package to be paid for	Click
	Can see the given payment code.	Verify Element Present

	Can choose an available bank.	Click
	Can see payment status	Verify Text Present
Edit Profile	Can press the edit profile button	Click
	Can delete profile data field	Clear Text
	Can fill in profile data fields	Set Text
	Can press save profile button	Click
	Can see notification that profile was successfully updated	Verify Element Present

Keywords are obtained from every condition that must be met from the features provided. The steps taken to fulfill these conditions have their respective keywords that must be carried out so that these steps can be carried out.

Test Cases are made on the Katalon Studio tools using the record and playback features. Record functions to record every action or keyword performed when using the website, while playback functions to play back every action that has been carried out through the record feature. The following is an example of a Test Case using the record and playback feature with the Keyword Driven Testing approach at Katalon Studio shown in Figure 2.

Item	Object	Input
1 - Open Browser		""
2 - Navigate To Url		"https://angkasa.rg.telko
3 - Click	a_Login	
4 - Set Text	input_Selamat datang ke	"reynaldi.octavially@gm
5 - Set Encrypted T	input_Selamat datang ke	"WyTIvxLuF62t86Fh13H
6 - Click	button_Login	
7 - Close Browser		

Fig. 2. Katalon Studio Test Case

- Test Suites Grouping

Test Suite is a grouping of tests based on Test Cases that have been created [16]. In Katalon studio, each Test Case can be grouped by page function or all Test Cases on the same page, so that each Test Case that has been made does not need to be executed one by one and does not require a long time. this grouping is called Test Suite as shown in Figure 3.

No.	ID
1	Test Cases/HomePage/About
2	Test Cases/HomePage/Contact
3	Test Cases/HomePage/Faq
4	Test Cases/HomePage/Learn
5	Test Cases/HomePage/Pricing
6	Test Cases/HomePage/Product

Fig. 3. Katalon Studio Test Suite

- Make Test Procedures

Specifies the order of execution of the entire test set. Test Procedures in this test are based on the user journey. In this test, they are grouped into three parts, namely for Angkasa Home Page, Dashboard, and Personal Information. This is adjusted to the page category provided by the website which is currently in the testing phase.

- Run Execution

Test Cases and Test Suites that have been created and sorted according to the Test Procedure will be executed using Katalon Studio. All Test Cases will be executed after being grouped into Test Suites to get a report from each Test Case created.

IV. RESULT AND DISCUSSION

After doing research using Keyword Driven Testing, the results and test analysis are as follows:

- Test Result

Based on the tests that have been carried out on the Angkasa Website, it was found that 10 keywords were obtained from Software Requirements Specification and obtained when testing using Katalon Studio. In accordance with the provisions of the ISO/IEC/IEEE 29119 document, keywords are divided into 2 levels, namely high and low. There are 4 keywords that fall into the high category, namely Open Browser, Close Browser, Navigate to Url, and Upload File. As for the low category, there are 6 keywords, namely Click, Set Text, Clear Text, Delay, Verify Element Present, and Verify Text Present. The keywords used can be seen in table II.

TABLE II. AUTOMATIC TEST RESULT

Keywords	Functions	Level
Open Browser	Open the browser in use.	
Close Browser	Close the browser in use.	High
Navigate to Url	Open the website address to be tested.	

Upload File	Upload and save photos on the website.	
Click	Press a button or link.	Low
Set Text	Fill in the field with text.	
Clear Text	Delete text in a field.	
Delay	Set the delay time while testing.	
Verify Element Present	Ensures elements appear on the page as they should.	
Verify Text Present	Make sure the text appears on the page as it should.	

Test Cases are made using the Record and Playback feature provided by Katalon Studio, this feature can make it easier for testers to test the Angkasa Website and automatically add keywords to the test document. But not all keywords can be added automatically, namely Upload, Verify Element Present and Verify Text Present. Upload is used to upload a file to the website, Verify Element Present serves to ensure an element such as images, text and others appear on a certain page as a sign that the test is successful and as it should be, while Verify Text Present is to ensure the text that appears on a particular page. These three features cannot be added automatically because Record and Playback only record what the user clicks or fills on the web page being tested, so if the command is not a click or fill in a field, it must be edited manually when the test is complete.

Tests are divided into 2 categories, namely successful and failed. The test is said to be successful if the input is correct and the results issued are also appropriate and the system declares the test successful. The test is said to have failed if firstly, the input is correct but the resulting output does not match, and secondly if the input is correct but the output of the system fails. The results of the tests that have been carried out on the Angkasa Website can be seen in table III.

TABLE III. AUTOMATIC TEST RESULT

Test Suites	**Number of Test Case**	**Status**	
		Passed	***Failed***
Angkasa Home Page	9	9	0
Angkasa Dashboard	11	11	0
Personal Information	3	3	0

The results of automatic testing using Katalon Studio show that all Test Suites are Passed and none are Failed, but different from the manual testing that has been done previously which had 1 failed from 3 Test Case on Test Suites on the Personal Information page. The results of manual testing can be seen in table IV.

TABLE IV. MANUAL TEST RESULT

Test Suites	**Number of Test Case**	**Status**	
		Passed	***Failed***
Angkasa Home Page	9	9	0
Angkasa Dashboard	11	11	0
Personal Information	3	2	1

- Test Analysis

In the test results, it is explained that there are some keywords that cannot be added automatically through the Record and Playback feature at the time of making the Test Case. This is because these keywords are used to ensure that an element appears on the page being tested, so keywords must be entered manually by selecting from a list of keywords provided by Katalon Studio.

By using the Keyword Driven Testing approach for making Test Cases on the Angkasa Website, all the tested features have the status passed and no features failed or failed. In this test there is a problem found by the researcher, namely the number of changes to the structure of the Angkasa Website, for example, such as id, name and also xpath so that when the researcher repeats the test on the same feature, the results will fail because there are changes to the website structure.

Testing by implementing a concept that is part of ISO/IEC/IEEE 29119-2 namely Keyword Driven Testing in The Test Design Process at Katalon Studio is a very appropriate step. Because Katalon Studio is a tool that supports testing using Keyword Driven Testing, because Katalon Studio has provided a list of keywords that can be used in the testing process, katalon Studio also provides Test Suites to group Test Cases based on their respective Page Functions, making the Test Case itself is divided into 3 options, namely Record and Payback, manual testing and testing using scripts. The results of automated testing using Katalon Studio for the test suites home page can be seen in Figure 4.

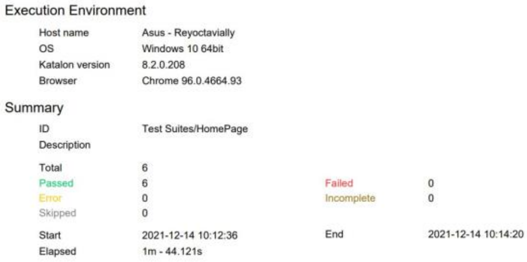


Fig. 4. Automated Testing Result

In the tests that have been carried out there are differences in results between automatic testing and manual testing. If an automatic test is carried out, all the results will get a PASSED status but if a manual test is carried out there is 1 feature that has a FAILED status, this happens because the FAILED status feature does not manage to display an image that matches the image uploaded by the user so it says FAILED is in the test. But the feature is PASSED when automatic testing is performed, that's because the automatic test only checks whether the image appears or not, does not check whether the image is damaged or not.

V. Conclusion

Keyword Driven Testing is very helpful in the software testing process because it is easy to add, remove and update if there are changes to the content of the website being tested. The identification of keywords to be used can be determined by exploration, business expert interviews, interface tests, and documented test procedures and test cases.

The keywords used in Keyword Driven Testing must comply with predetermined procedures, which must contain a verb, have an imperative form, provide a description and others.

In testing a software, two tests must be carried out, namely manual testing and automatic testing, because the two tests have different results, so it can be concluded that the two tests cannot be separated. Although automatic testing is easier and faster, manual testing can be more thorough and have more accurate results.

In this study, automated testing analysis with Keyword Driven Testing was carried out on the Angkasa Website which is a cloud-based website being developed by Telkom University, testing was carried out using the Katalon Studio tools that fully support Keyword Driven Testing. This study also uses the concept of ISO/IEC/IEEE 29119-2, namely Keyword Driven Testing in The Test Design Process. The conclusions obtained from this research are:

- Automatic Testing successfully applied 10 keywords that were used to test 26 features on the Angkasa Website with PASSED results for all of its features.
- Manual Testing successfully applied 10 keywords that were used to test 26 features on the Angkasa Website with PASS results for all 25 features and FAILED for 1 feature.
- Keyword Driven Testing in The Test Design Process has been successfully implemented on the Katalon Studio tools by using the Record and Payback features to make it easier to search, add, and use the required keywords. Katalon Studio provides a list of keywords that are ready to be used for the Record and Playback feature, manual testing or testing using scripts.
- The keywords that Katalon Studio can't read automatically are because they don't give any action but only as a marker if an element is successfully displayed and indicates a successful test.

Execution between Test Cases requires a time lag to avoid arbitrary results.

Acknowledgement

This work was part of matching fund Kedaulatan Indonesia dalam Reka Cipta (Kedaireka) – Kampus Merdeka supported by Kementrian Pendidikan dan Kebudayaan Republik Indonesia, Jakarta 10270.

References

[1] E. Pelivani and B. Cico, "A comparative study of automation testing tools for web applications," Jun. 2021. doi: 10.1109/MECO52532.2021.9460242.
[2] X. Meng, "Analysis of Software Automation Test Protocol," 2011.
[3] SKR Engineering College, Institute of Electrical and Electronics Engineers. Madras Section, and Institute of Electrical and Electronics Engineers, International Conference on Energy, Communication, Data Analytics & Soft Computing (ICECDS) - 2017 : 1st & 2nd August 2017.
[4] United International College, IEEE Computer Society, and Institute of Electrical and Electronics Engineers, Proceedings of 2015 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) : date and venue: 10-12 December 2015, United International College, Zhuhai, China.
[5] Jouf University and Institute of Electrical and Electronics Engineers, 2019 International Conference on Computer and Information Sciences (ICCIS) : Jouf University - Aljouf - kingdom of Saudi Arabia, 03 - 04 April 2019.
[6] Jingfan Tang, Xiaohua Cao, and Albert Ma, *Proceedings of the IEEE International Conference on Automation and Logistics : Qingdao, China September 2008.* IEEE Xplore, 2008.
[7] U. Nugraha, S. Atikah Nurduha Robaiah, and D. Rospinoedji, "Testing The Information System Software Using Behavior Driven Development Method," *Palarch's Journal Of Archaeology Of Egypt*, vol. 17, no. 10, 2020.

[8] R. Gupta and N. Bajpai, “A keyword-driven tool for testing web applications (KeyDriver),” *IEEE Potentials*, vol. 33, no. 5, pp. 35–42, Sep. 2014, doi: 10.1109/MPOT.2012.2202135.

[9] R. Rwemalika, M. Kintis, M. Papadakis, Y. le Traon, and P. Lorrach, “On the evolution of keyword-driven test suites,” in *Proceedings - 2019 IEEE 12th International Conference on Software Testing, Verification and Validation, ICST 2019*, Apr. 2019, pp. 335–345. doi: 10.1109/ICST.2019.00040.

[10] N. Srivastava, U. Kumar, P. Singh, and P. Singh, “Software and Performance Testing Tools,” *Journal of Infor-matics Electrical and Electronics Engineering*, vol. 02, no. 001, pp. 1–12, 2021.

[11] Institute of Electrical and Electronics Engineers., *2012 7th International Workshop on Automation of Software Test (AST) : proceedings : June 2-3, 2012, Zurich, Switzerland.* IEEE, 2012.

[12] I. Computer Society and S. C. Engineering Standards Committee, “Software and systems engineering-Software testing-Part 3: Test documentation Ingénierie du logiciel et des systèmes-Essais du logiciel-Partie 3: Documentation des essais,” 2013.

[13] International Electrotechnical Commission, Institute of Electrical and Electronics Engineers, and International Electrotechnical Commission. Technical Committee 91, *Standard for automatic test markup language (ATML) test adapter description*.

[14] B. Bhagat, S. Bhattacharjee, and S. Ratre, “Software Testing Techniques & Automation Tools,” *Mukt Shabd J.*, vol. 9, no. 5, pp. 5957–5962, 2020, doi: 09.0014.MSJ.2020.V9I5.0086781.104235.

[15] M. Mirzaaghaei, F. Pastore, and M. Pezzè, “Supporting Test Suite Evolution through Test Case Adaptation,” 2012, doi: 10.1109/ICST.2012.35.

[16] Ahlam S. A. Ansari, Kailas K. Devadkar, and Prachi Gharpure, “Optimization of test suite-test case in regression test,” 2013.

AUTHOR GUIDELINES

1. Manuscript criteria

- The article has never been published or in the submission process on other publications.
- Submitted articles could be original research articles or technical notes.
- The similarity score from plagiarism checker software such as Turnitin is 20% maximum.
- For December 2021 publication onwards, Ultimatics : Jurnal Teknik Informatika will be receiving and publishing manuscripts written in English only.

2. Manuscript format

- Article been type in Microsoft Word version 2007 or later.
- Article been typed with 1 line spacing on an A4 paper size (21 cm x 29,7 cm), top-left margin are 3 cm and bottom-right margin are 2 cm, and Times New Roman's font type.
- Article should be prepared according to the following author guidelines in this template. Article contain of minimum 3500 words.
- References contain of minimum 15 references (primary references) from reputable journals/conferences

3. Organization of submitted article

The organization of the submitted article consists of Title, Abstract, Index Terms, Introduction, Method, Result and Discussion, Conclusion, Appendix (if any), Acknowledgment (if any), and References.

- Title
  The maximum words count on the title is 12 words (including the subtitle if available)
- Abstract
  Abstract consists of 150-250 words. The abstract should contain logical argumentation of the research taken, problem-solving methodology, research results, and a brief conclusion.
- Index terms
  A list in alphabetical order in between 4 to 6 words or short phrases separated by a semicolon (;), excluding words used in the title and chosen carefully to reflect the precise content of the paper.
- Introduction
  Introduction commonly contains the background, purpose of the research, problem identification, research methodology, and state of the art conducted by the authors which describe implicitly.
- Method
  Include sufficient details for the work to be repeated. Where specific equipment and materials are named, the manufacturer's details (name, city and country) should be given so that readers can trace specifications by contacting the manufacturer. Where commercially available software has been used, details of the supplier should be given in brackets or the reference given in full in the reference list.
- Results and Discussion
  State the results of experimental or modeling work, drawing attention to important details in tables and figures, and discuss them intensively by comparing and/or citing other references.
- Conclusion
  Explicitly describes the research's results been taken. Future works or suggestion could be explained after it
- Appendix and acknowledgment, if available, could be placed after Conclusion.
- All citations in the article should be written on References consecutively based on its' appearance order in the article using Mendeley (recommendation). The typing format will be in the same format as the IEEE journals and transaction format.

4. Reviewing of Manuscripts

Every submitted paper is independently and blindly reviewed by at least two peer-reviewers. The decision for publication, amendment, or rejection is based upon their reports/recommendations. If two or more reviewers consider a manuscript unsuitable for publication in this journal, a statement explaining the basis for the decision will be sent to the authors within six months of the submission date.

5. Revision of Manuscripts

Manuscripts sent back to the authors for revision should be returned to the editor without delay (maximum of two weeks). Revised manuscripts can be sent to the editorial office through the same online system. Revised manuscripts returned later than one month will be considered as new submissions.

6. Editing References

- **Periodicals**
  J.K. Author, “Name of paper,” Abbrev. Title of Periodical, vol. x, no. x, pp. xxx-xxx, Sept. 2013.

- **Book**
  J.K. Author, “Title of chapter in the book,” in Title of His Published Book, xth ed. City of Publisher, Country or Nation: Abbrev. Of Publisher, year, ch. x, sec. x, pp xxx-xxx.

- **Report**
  J.K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Rep. xxx, year.

- **Handbook**
  Name of Manual/ Handbook, x ed., Abbrev. Name of Co., City of Co., Abbrev. State, year, pp. xxx-xxx.

- **Published Conference Proceedings**
  J.K. Author, “Title of paper,” in Unabbreviated Name of Conf., City of Conf., Abbrev. State (if given), year, pp. xxx-xxx.

- **Papers Presented at Conferences**
  J.K. Author, “Title of paper,” presented at the Unabbrev. Name of Conf., City of Conf., Abbrev. State, year.

- **Patents**
  J.K. Author, “Title of patent,” US. Patent xxxxxxx, Abbrev. 01 January 2014.

- **Theses and Dissertations**
  J.K. Author, “Title of thesis,” M.Sc. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year. J.K. Author, “Title of dissertation,” Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

- **Unpublished**
  J.K. Author, “Title of paper,” unpublished. J.K. Author, “Title of paper,” Abbrev. Title of Journal, in press.

- **On-line Sources**
  J.K. Author. (year, month day). Title (edition) [Type of medium]. Available: http:// www.(URL) J.K. Author. (year, month). Title. Journal [Type of medium]. volume(issue), pp. if given. Available: http://www.(URL) Note: type of medium could be online media, CD-ROM, USB, etc.

7. Editorial Adress

Universitas Multimedia Nusantara
Jl. Scientia Boulevard, Gading Serpong
Tangerang, Banten, 15811
Email: **ultimatics@umn.ac.id**

Paper Title

Subtitle (if needed)

Author 1 Name[1], Author 2 Name[2], Author 3 Name[2]

[1] Line 1 (of affiliation): dept. name of organization, organization name, City, Country
Line 2: e-mail address if desired

[2] Line 1 (of affiliation): dept. name of organization, organization name, City, Country
Line 2: e-mail address if desired

Accepted on mmmmm dd, yyyy
Approved on mmmmm dd, yyyy

***Abstract*—This electronic document is a "live" template which you can use on preparing your ULTIMATICS paper. Use this document as a template if you are using Microsoft Word 2007 or later. Otherwise, use this document as an instruction set. Do not use symbol, special characters, or Math in Paper Title and Abstract. Do not cite references in the abstract.**

***Index Terms*—enter key words or phrases in alphabetical order, separated by semicolon (;)**

I. Introduction

This template, modified in MS Word 2007 and saved as a Word 97-2003 document, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. Margins, column widths, line spacing, and type styles are built-in here. The authors must make sure that their paper has fulfilled all the formatting stated here.

Introduction commonly contains the background, purpose of the research, problem identification, and research methodology conducted by the authors which been describe implicitly. Except for Introduction and Conclusion, other chapter's title must be explicitly represent the content of the chapter.

II. Ease of Use

A. Selecting a Template

First, confirm that you have the correct template for your paper size. This template is for ULTIMATICS. It has been tailored for output on the A4 paper size.

B. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not add any kind of pagination anywhere in the paper. Please take note of the following items when proofreading spelling and grammar.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units (SI units are encouraged).
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter," not "webers/m2." Spell units when they appear in text: "...a few henries," not "...a few H."
- Use a zero before decimal points: "0.25," not ".25."

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number the equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop.

$$\int_0^{r_2} F(r,\phi)\, dr\, d\phi = [\sigma r_2 / (2\mu_0)] \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ..."

D. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o."
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset," not an "insert." The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively."
- In your paper title, if the words "that uses" can accurately replace the word using, capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect," "complement" and "compliment," "discreet" and "discrete," "principal" and "principle."
- Do not confuse "imply" and "infer."
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al."
- The abbreviation "i.e." means "that is," and the abbreviation "e.g." means "for example."

IV. USING THE TEMPLATE

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention as below

ULTIMATICS_firstAuthorName_paperTitle.

In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper. Please take note on the following items.

A. Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is "Heading 5."

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles, named "Heading 1," "Heading 2," "Heading 3," and "Heading 4", are prescribed.

C. Figures and Tables

Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1," even at the beginning of a sentence.

TABLE I. TABLE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy		

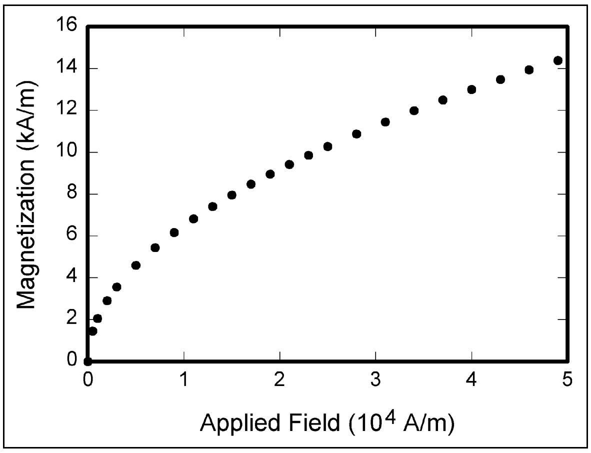


Fig. 1. Example of a figure caption

V. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Appendix

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank ... ." Instead, write "F. A. Author thanks ... ." You could also state the sponsor and financial support acknowledgments here.

References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

ISSN 2085-4552

PRESS

Universitas Multimedia Nusantara
Scientia Garden Jl. Boulevard Gading Serpong, Tangerang
Telp. (021) 5422 0808 | Fax. (021) 5422 0800