UNIVERSITAS UDAYANA

ISSN: 2301-5373
E-ISSN: 2654-5101
Volume 10 • Number 3 • February 2022

JELIKU

Jurnal Elektronik Ilmu Komputer Udayana

Informatics Study Program
Faculty of Mathematics and Natural Sciences
Udayana University

Table of Contents

SUSUNAN DEWAN REDAKSI

JURNAL ELEKTRONIK ILMU KOMPUTER UDAYANA (JELIKU)

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Vocal Tone Precision Detection Using Harmonic Product Spectrum And K-Nearest Neighbor Classification

Made Sri Ayu Apsari[a1], I Made Widiartha[a2], Made Agung Raharja[a3], I Gede Santi Astawa[a4], I Gede Arta Wibawa[a5], Ida Bagus Made Mahendra[a6]

[a]Informatics Department, Faculty of Mathematics and Natural Sciences, University of Udayana
South Kuta, Badung, Bali, Indonesia
[1]ayuapsaarii25@gmail.com
[2]madewidiartha@unud.ac.id
[3]made.agung@unud.ac.id
[4]santi.astawa@unud.ac.id
[5]gede.arta@unud.ac.id
[6]ibm.mahendra@unud.ac.id

Abstract

The progress of the digital era that is happening today, encourages rapid development in technology and science, one of which is in the field of art. Of all performing arts, the art of singing is the most complex, which requires a lot of preparation and practice. Everyone has a different type of voice. Males generally have three types of voice, namely bass, baritone, and tenor, while women generally have three types of voice, namely contralto (alto), mezzo-soprano, and soprano. However, not everyone knows what kind of voice they have. Therefore, this study will focus on classifying the human voice. In this study, the author uses the Harmonic Product Spectrum (HPS) and K-Nearest Neighbor (K-NN) algorithms. The data used is in the form of primary voice recording data obtained from 258 participants (male and female), where each person has 8 sound files, namely do, re, mi, fa, sol, la, si, and do'. saved in .wav format. From the research conducted, the test was carried out using the K-NN and K-NN methods with Hyperparameters. The results obtained in the form of accuracy of 74% and 81%, so that the Harmonic Product Spectrum (HPS) and K-Nearest Neighbor (K-NN) algorithms give good results for determining the type of human voice.

Keywords : *Harmonic Product Spectrum, K-Nearest Neighbor, vocal tone detection*, *vocal type classification*, *vocal range.*

1. Introduction

The progress of the digital era that is happening today, encourages rapid development of technology and science. Technological developments have changed various aspects of human life such as social, health, economic, and arts and culture aspects. Aspects of art and culture are changing quite rapidly due to technological developments, especially in the field of music. Music is a beautiful sound that can be heard. Sounds in music can be sourced from tools that are able to make sounds and are sourced from humans[11]

At first music could only be heard directly when the music was played, but with advances in technology, music can be heard at any time with the help of recordings. The emergence of various songs that match the interests of the community makes people creative by singing back songs that are on the rise using their own characteristics but still maintaining the originality of the song.

Of all performing arts, the art of singing is the most complex, which requires a lot of preparation and practice[8]. Everyone has a different type of voice where each type of voice has a different vocal range. Males generally have three types of voices, namely bass, baritone, and tenor, while women generally have three types of voices, namely contralto (alto), mezzo-soprano, and soprano.

The way that can be done to find out the type of voice you have is to compare your voice with a musical instrument, generally the musical instrument used is the piano[7]. In addition, the way that can be done to find out the type of voice is to do vocal lessons where a vocal coach will help find out the type of voice you have[10].

The suitability of the type of voice and tone with the song is important because if someone ignores this, the voice when singing seems discordant or false. This happens because not everyone who wants to know the type of voice has a musical instrument or goes through vocal lessons so that it becomes an obstacle when singing. With this phenomenon, we need a technology that can help detect the type of person's voice without having to use a musical instrument or visit a vocal coach.

Detection of musical instrument tones is also carried out on angklung musical instruments where the system used obtains optimal results with an accuracy rate of 88.78%[3]. The system with the Harmonic Product Spectrum (HPS) algorithm is used to see the basic frequency contained in the input signal then the K-Nearest Neighbor (KNN) classification method is used to detect and recognize the tone that is being played. The android-based guitar tuner application uses the Fast Fourier Transform (FFT) algorithm and the Harmonic Product Spectrum (HPS) algorithm.

The guitar tuner application works well and can help in the guitar tuning process[1]. In the study of chord and melody detection in fingerstyle wav files using the DWPT and K-NN methods, the results obtained are 99.07% for single chord detection, 100% for single note detection, and 83.11% average accuracy for detecting 40 fingerstyle music. In this study, the K-Nearest Neighbor (KNN) method was used as a classification method. The tones used as training data were 355 chord recording data and 125 single tone recording data. The data tested in this study were 195 chord recordings, 75 single note recordings, and 8 fingerstyle music, each of which was recorded 5 times.

The Harmonic Product Spectrum (HPS) algorithm is proven to be able to be used to determine the basic frequency of the tone and the K-Nearest Neighbor (KNN) classification can be used to classify sound types based on tone. In the study of music classification based on active frequency using the K-Nearest Neighbor (KNN) method can classify music with several different genres. The level of accuracy obtained in this study is 70%. Pitch detection using the Harmonic Product Spectrum (HPS) algorithm results in sharper harmonics in the spectrum[4].

Based on the explanation of the problems above, the author intends to conduct research to determine the accuracy of vocal tone detection using the Harmonic Product Spectrum (HPS) algorithm and determine the type of sound using the K-Nearest Neighbor (KNN) classification. The process carried out to detect vocal tones requires sound signal processing. so as to give the appropriate results.

2. Research Methods

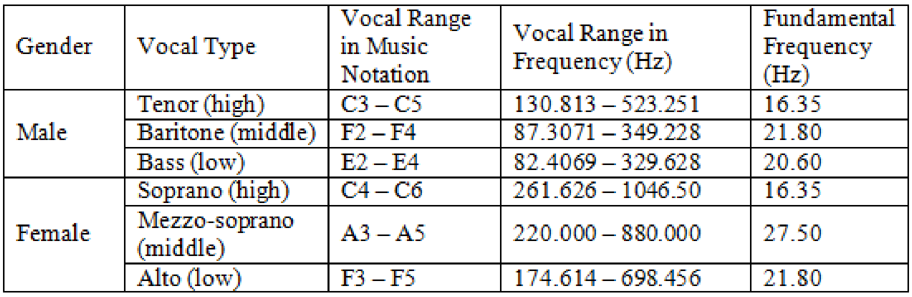
2.1 Tone

A tone is a sound whose frequency has been set. IMC (International Music Council) has set the tone a' = 440, which means that the a' tone must vibrate 440 vibrations in 1 second, or in other words the a' tone = 440 vibrations per second[11].

2.2 Voice Type

There are four main types of voices in humans when singing, soprano and alto (contralto) for women, and tenor and bass for men. However, for each type there are subtypes such as mezzo-soprano for women, and baritone for In general, soprano and tenor voices have a higher range than alto and bass voices. Soprano is the highest voice type in women, with the most common subtypes being lyric and mezzo. Both can sing the same range, but lyric soprano has a lighter tone, while mezzo notes have a deeper tone. The lowest voice in women is (alto) contralto which is divided into two parts, namely the first and second alto. The first alto has a lighter tone, while the second alto is heavier[2].

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gender	Vocal Type	Vocal Range in Music Notation	Vocal Range in Frequency (Hz)	Fundamental Frequency (Hz)
Male	Tenor (high)	C3 – C5	130.813 – 523.251	16.35
	Baritone (middle)	F2 – F4	87.3071 – 349.228	21.80
	Bass (low)	E2 – E4	82.4069 – 329.628	20.60
Female	Soprano (high)	C4 – C6	261.626 – 1046.50	16.35
	Mezzo-soprano (middle)	A3 – A5	220.000 – 880.000	27.50
	Alto (low)	F3 – F5	174.614 – 698.456	21.80

Figure 1 Voice Type

2.3 Frequency

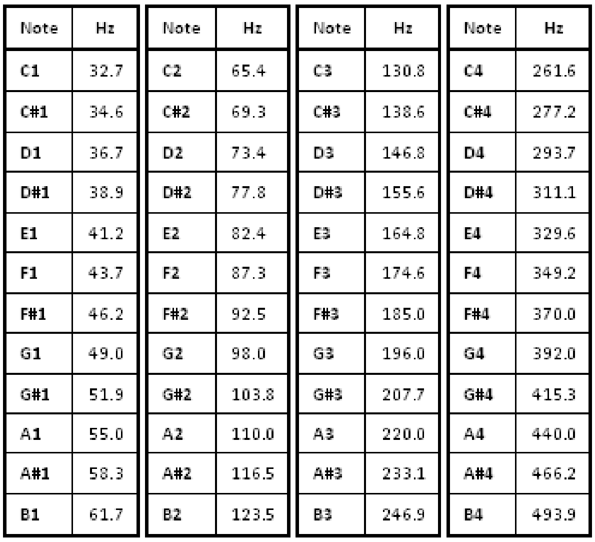
Sound frequency is the number of vibrations or the number of vibrations that occur per second in a sound wave or sound wave. Frequency is also defined as the number of changes in pressure per second or frequency per second in units of cycles per second (cls) or Hertz (Hz). The nature of sound is determined by its frequency and intensity.

Based on the frequency, sound or voice is divided into three frequency regions, namely:

a. Sound frequency between 0 – 20 Hz (Infrasonic Area) This sound frequency, for example, is ground vibration, building vibration and car truck.

b. Sound frequency between 20 – 20,000 Hz (Sonic Frequency/hearing) Ear sensitivity dB = 0 occurs at a frequency of 1000 Hz, where the international average threshold value lies in the 1000 Hz area.

c. Sound frequency above 20,000 Hz (Ultrasonic Area) 16 In the medical field, this frequency functions in determining 3 things, namely: treatment, destruction/destructive and diagnosis. This is because the high frequency has a fairly large network penetration power.



Note	Hz	Note	Hz	Note	Hz	Note	Hz
C1	32.7	C2	65.4	C3	130.8	C4	261.6
C#1	34.6	C#2	69.3	C#3	138.6	C#4	277.2
D1	36.7	D2	73.4	D3	146.8	D4	293.7
D#1	38.9	D#2	77.8	D#3	155.6	D#4	311.1
E1	41.2	E2	82.4	E3	164.8	E4	329.6
F1	43.7	F2	87.3	F3	174.6	F4	349.2
F#1	46.2	F#2	92.5	F#3	185.0	F#4	370.0
G1	49.0	G2	98.0	G3	196.0	G4	392.0
G#1	51.9	G#2	103.8	G#3	207.7	G#4	415.3
A1	55.0	A2	110.0	A3	220.0	A4	440.0
A#1	58.3	A#2	116.5	A#3	233.1	A#4	466.2
B1	61.7	B2	123.5	B3	246.9	B4	493.9

Figure 2 Grand Piano Frequency

2.4 *Fast Fourier Transform (FFT)*

Fast Fourier Transform (FFT) is a Fourier transform algorithm developed from the Discrete Fourier Transform (DFT) algorithm. The Fast Fourier Transform algorithm is very efficient in calculating the DFT coefficient and can reduce the enormous computational complexity. FFT is a method that converts signals from time domain to frequency domain. By using this FFT method, the computation rate of the Fourier transform calculation can be increased. The formula of the FFT method can be defined as follows in equation (1).

$X[k]$ = $\sum Nn$=-11 $x(n)WNkn$ (1)

2.5 *Harmonic Product Spectrum (HPS)*

Harmonic Product Spectrum (HPS) algorithm is an algorithm to determine the tone in the frequency domain[6]. This algorithm is also a pitch detection algorithm based on Fourier transform. This algorithm also takes advantage of the tendency of pitched musical signals to show strong harmonic structures. The algorithm works by down-sampling spectra and multiplying spectra[9]. Fast Fourier Transform (FFT) algorithm is used to represent signals in spectral form. If the input signal is a musical note, then its spectrum must consist of a series of peaks corresponding to the fundamental frequency with harmonic components. The downsampling process is the process of compressing the spectrum to get the fundamental frequency of the signal.

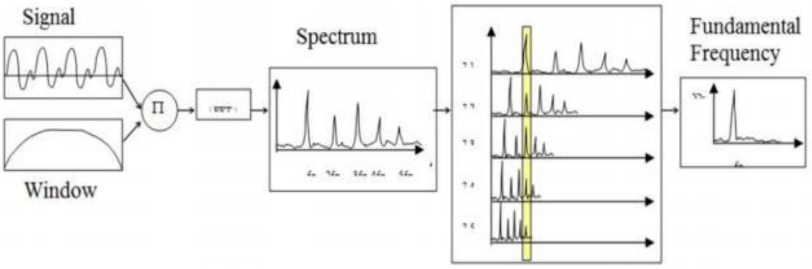
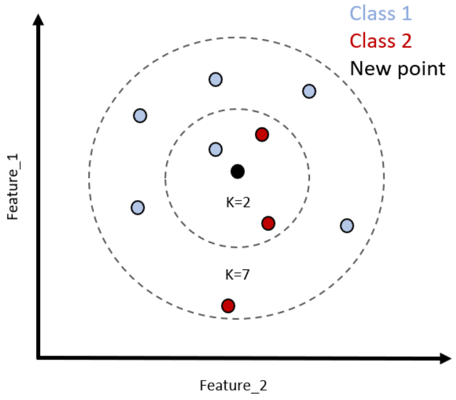


Figure 3 Harmonic Product Spectrum

2.6 *K-Nearest Neighbor (K-NN)*

K-Nearest Neighbor (KNN) algorithm is a method for classifying objects based on the learning data that is closest to the object[5]. This method is widely used in the field of pattern recognition. K-Nearest Neighbor classification is based on comparing the given test data with similar training data. The training data is described by n attributes. Each data represents a point in the n-dimensional space. In this way, all training data will be stored in the n-dimensional pattern space. When given unrecognized data, the K-Nearest Neighbor classification will look for a pattern space for the k training data closest to the unknown test data. This method has a way of working that is to find the closest distance from the value to be evaluated with its nearest neighbor in a data. At the classification stage, the same features will be calculated to perform data testing. The distance from the new vector to the entire training sample vector will be calculated and the closest number of k pieces will be taken.



Produksi Pribadi Penulis, Stephen Fordham

Figure 4 K-Nearest Neighbor

2.7 Python

Python is a multipurpose interpretive programming language. Unlike other languages which are difficult to read and understand, python places more emphasis on code readability to make it easier to understand syntax. This makes Python very easy to learn both for beginners and for those who have mastered other programming languages.

This language first appeared in 1991, designed by a person named Guido van Rossum. Until now Python is still being developed by the Python Software Foundation. Python language supports almost all operating systems, even for the Linux operating system, almost all distributions already include Python in it.

2.8 Data

The data used in this study is primary data. Data was obtained by recording the tones of 258 participants (male and female) recorded in .wav format. The audio format .wav or waveform is used because the sound data in this format has not been compressed so that it can be said that the waveform file is a raw file or pure data from recorded sound. In this study, the tone used is a single basic tone. Each note that is sung is a single basic note (do, re, mi, fa, sol, la, si, and high do). One participant has 8 files that will be used as system input. The training data and test data used in this study were divided into 70% and 30% of the total data.

2.9 Flowchart

The inputted data will go through several stages. The recording file is a file in .wav format. Then it will enter the preprocessing stage to equalize the input signal. At this stage consists of windowing and FFT. Then it converts the signal in the time domain into the frequency domain. Harmonic Product Spectrum (HPS) algorithm is then used to remove harmonic frequencies from the sound signal. Then the classification stage uses the K-Nearest Neighbor (KNN) method to determine the type of sound.

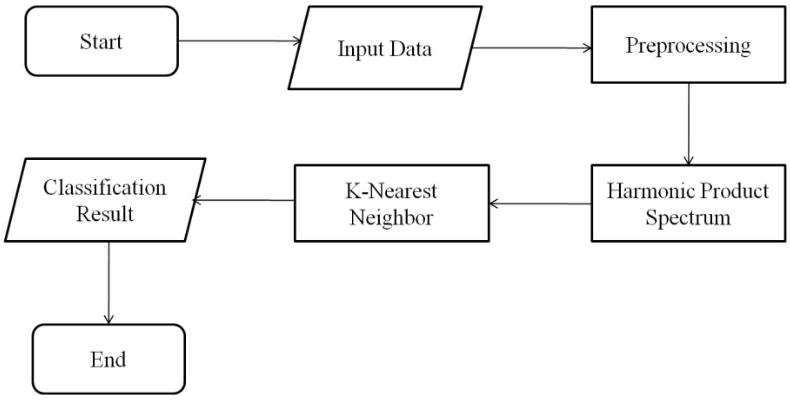
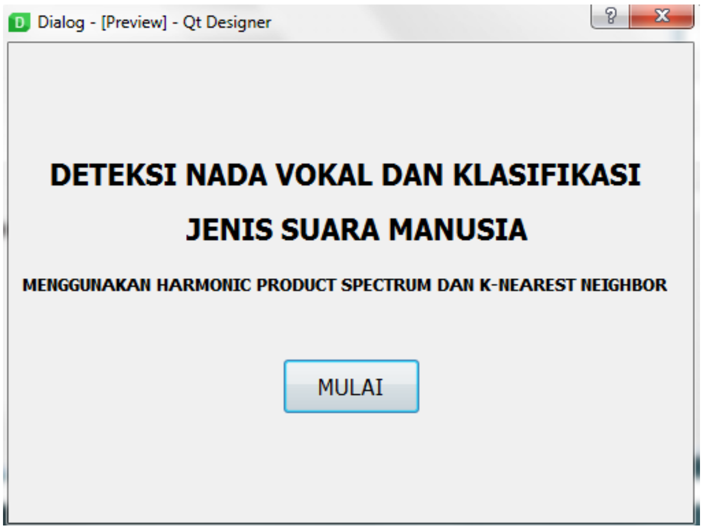


Figure 5 Flowchart System

2.10 System Design

At this stage the system created will be represented by a design or description including interface representation and coding procedures. This stage is also included in the representation of the appearance of the software that will be made. The system interface used in this study uses the desktop-based Python programming language using the PyQt5 framework. The system created has two pages, namely the front page and the classification page.



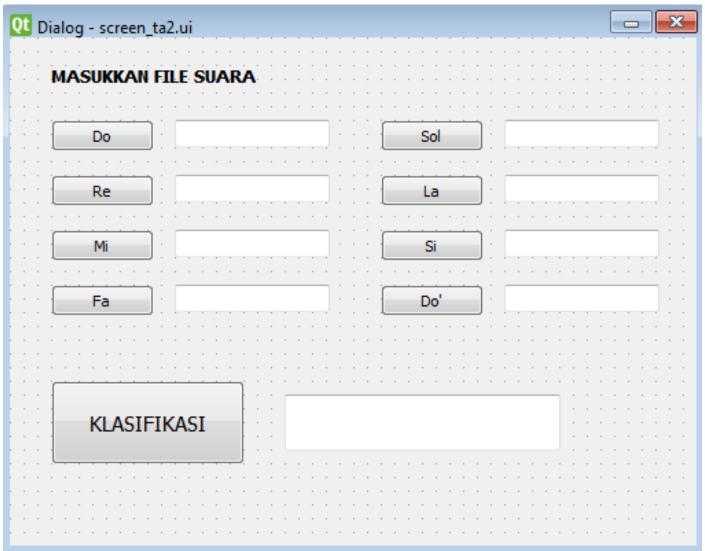
Dialog - [Preview] - Qt Designer

DETEKSI NADA VOKAL DAN KLASIFIKASI JENIS SUARA MANUSIA

MENGGUNAKAN HARMONIC PRODUCT SPECTRUM DAN K-NEAREST NEIGHBOR

Figure 6 System GUI

Figure 6 is the front page of this system. This page is the initial display when the user starts the system. On this display there is a label and a “START” button. When the “START” button is clicked, the user will be directed to the next page, namely the classification page.



Dialog - screen_ta2.ui

MASUKKAN FILE SUARA

Figure 7 System GUI (1)

Figure 7 is this system’s classification page. This classification page is used to classify sound types based on files that have been inputted by the user. On this page there are buttons used to enter sound files, namely “Do”, “Re”, “Mi”, “Fa”, “Sol”, “La”, “Si”, and “Do'”. Then there is also the "CLASSIFICATION" button which is used to carry out the process of classifying the type of sound. The results obtained will be displayed in LineEdit.

3. Result and Discussion

3.1 Analysis of the effect of the value of k on accuracy.

The value of k in the K-Nearest Neighbor classification is the number of nearest neighbors used to determine the class of a data. The optimal value of k will increase the accuracy of the system. k = 2 to k = 23 will find the most optimal value to be used in the system. The amount of training data used is 180 data and the amount of test data used is 77 data with features, namely do, re, mi, fa, sol, la, si and do'.

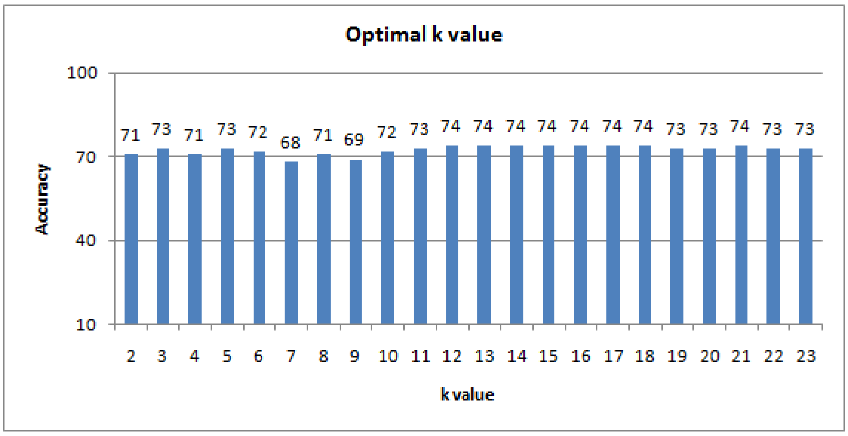


Figure 8 Optimal k Value

From figure 8, it can be seen that the k value with the highest accuracy lies in the k value between 12 to 22 and is at a stable accuracy of 73% and 74%.

3.2 Analysis of the effect of the value of k using hyperparametertuning on accuracy.

The value of k in the K-Nearest Neighbor classification is the number of nearest neighbors used to determine the class of a data. The optimal value of k will increase the accuracy of the system. k = 2 to k = 23 will find the most optimal value to be used in the system. The amount of training data used is 180 data and the amount of test data used is 77 data with features, namely do, re, mi, fa, sol, la, si and do'. The optimal k value in the second way is with the K-NN Hyperparameter, where this Hyperparameter serves to increase the accuracy of the system to be made. The parameters used are leaf_size, p, and n_neighbor.

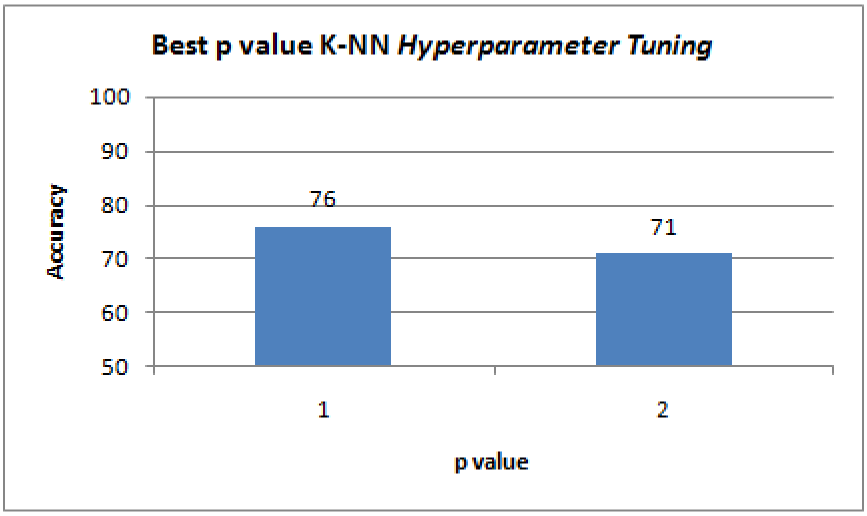


Figure 9 The best p value of K-NN Hyperparameter Tuning

From figure 9 we can determine that the best p value is p = 1 with an accuracy of 76%.

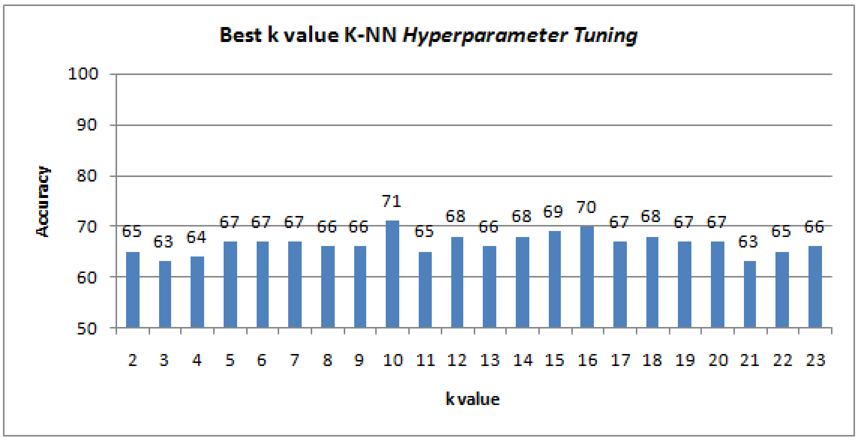


Figure 10 Best k value K-NN Hyperparameter Tuning

From figure 10 we can determine that the best k value is k=10 with an accuracy of 71%.

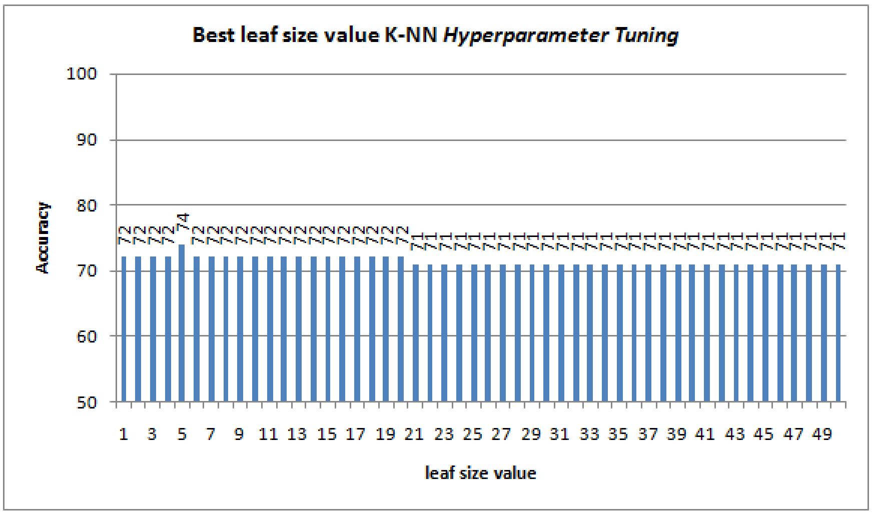
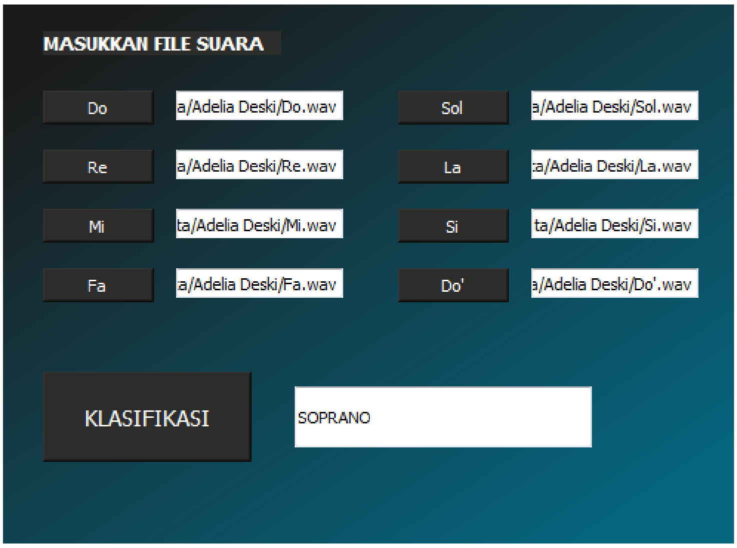


Figure 11 Best leafsize value for K-NN Hyperparameter Tuning

From figure 11, it can be seen that the best leafsize value is leafsize = 6 with an accuracy of 74%.

The three parameters are searched for the best for the system by combining all the possible parameters used. The results obtained from K-NN using this hyperparameter are leaf_size=6, p=1, and n_neighbors=10. The K-NN model uses Hyperparameter Tuning with parameters leaf_size=6, p=1, and n_neighbors=10 resulting in an accuracy of 81%, so this K-NN model will be used in the system to classify the type of human voice.

Based on the analysis carried out, it can be seen that the K-NN model with Hyperparameter Tuning with parameters leaf_size=6, p=1, and n_neighbors=10 each has the highest accuracy. The test will be carried out using a confusion matrix. The test was carried out using a dataset of 258 audio files with the distribution of the training data as much as 180 data and the number of testing data as 77 data. This test is carried out using the parameters leaf_size=6, p=1, and n_neighbors=10.



a/Adelia Deski/Do.wav
a/Adelia Deski/Re.wav
ta/Adelia Deski/Mi.wav
a/Adelia Deski/Fa.wav
a/Adelia Deski/Sol.wav
:a/Adelia Deski/La.wav
ta/Adelia Deski/Si.wav
a/Adelia Deski/Do'.wav

SOPRANO

Figure 12 System Result

Figure 12 is a classification page display with the files that have been entered and the results that have been displayed. In the picture, it can be seen that the result of the classification is SOPRANO, which means that the person has a SOPRANO type of voice.

4. Conclusion

From the implementation of the research that has been done and the results that have been obtained, the following conclusions can be drawn.
1. From the results of the research conducted, the K-Nearest Neighbor method can be used to classify the types of human voice well.
2. The accuracy obtained in classifying the type of sound using K-Nearest Neighbor is 74% with an optimal k value of k above 12 and the accuracy of K-Nearest Neighbor classification using Hyperparameter Tuning is 81% with the parameter used, namely leaf_size=6 , p=1, and n_neighbors=10.
3. The composition and amount of training data used greatly affect the accuracy of the system, with the amount of valid training data increasing the tendency of accuracy.

References

[1] F. Abdillah, "Implementasi Algoritma *Fast Fourier Transform (FFT)* dan Algoritma *Harmonic Product Spectrum (HPS)* pada Tuner Gitar Berbasis Android," *Jurnal Nuansa Informatika,* vol.11, no.2, p.18-25, 2017.

[2] P.S. Phillips, Singing For Dummies, 3rd Edition, United States: Wiley, 2021, ch.1, pp. 15-25.

[3] R. Ashary, R. Patmasari and S. Saidah, "Sistem Deteksi Nada Alat Musik Angklung Menggunakan Metode *Harmonic Product Spectrum*," *e-Proceeding of Engineering,* vol.6, no.1, p.1039-1046, 2019.

[4] M.I. Fauzi, R. Magdalena and B. Hidayat, "Deteksi Akor dan Melodi pada File Wav Gitar *Fingerstyle* Menggunakan Metode DWPT Dan K-NN," *JESCE : Journal of Electrical and System Control Engineering,* vol.3, no.2, p.116-125, 2020.

[5] C.C. Aggarwal and S. Sathe, Outliner Ensembles: An Introduction, New York City: Springer International Publishing, 2017, ch.6, pp. 214.

[6] A.R. Jayan, Speech and Audio Signal Procesing, India: PHI Learning Pvt. Ltd., 2017, pp. 53.

[7] E. Lutters, Kunci Sukses Menjadi Aktor, Jakarta: Gramedia Widiasarana Indonesia, 2018, ch.2, pp.107.

[8] E.T.L.Caruso, E.Tetrazzini and L.Caruso, The Art of Singing, Germany: Outlook Verlag, 3rd Edition, 2018 pp.35.

[9] D. Zhang and K. Wu, Pathological Voice Analysis, Singapore: Springer Singapore, 2020, ch.3, pp. 49.

[10] R. Bale, Teaching with Confidence in Higher Education: Applying Strategies from the Performing Arts, United Kingdom: Taylor & Francis, 2020, ch.4.

[11] N. Simanungkalit, Teknik Vokal Paduan Suara, Jakarta: Gramedia Pustaka Utama, 2013, ch.1, pp.1-6.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Music Genre Classification Using Modified K-Nearest Neighbor (MK-NN)

I Nyoman Yusha Tresnatama Giri[a1], Luh Arida Ayu Rahning Putri[a2], Gst. Ayu Vida Mastrika Giri[a3], I Gusti Ngurah Anom Cahyadi Putra[a4], I Made Widiartha[a5], I Wayan Supriana[a6]

[a]Informatics Department, Faculty of Mathematics and Natural Sciences, University of Udayana
South Kuta, Badung, Bali, Indonesia
[1]yusatresnatama11@gmail.com
[2]rahningputri@unud.ac.id
[3]vida.mastrika@cs.unud.ac.id
[4]anom.cp@unud.ac.id
[5]madewidiartha@unud.ac.id
[6]wayan.supriana@unud.ac.id

Abstract

The genre of music is a grouping of music according to their resemblance to one another and commonly used to organize digital music. To classify music into certain genres, one can do it by listening to the music one by one manually, which will take a long time so that automatic genre assignment is needed which can be done by a number of methods, one of which is the Modified K-Nearest Neighbor.

Modified K-Nearest Neighbor method is a further development of its former method called K-Nearest Neighbor method which adds several additional processes such as validity calculations and weight calculations to provide more information in the selection class for the testing data.

Research to find the best H value shows that the H = 70% of the training data is able to produce an accuracy of 54.100% with K = 5 and the proportion ratio of test data and training data is 20:80 (fold 5). The best H value is then used for further testing, which is to compare the K-Nearest Neighbor method with the Modified K-Nearest Neighbor method using two different proportions of test data and training data and each proportion of data also tests a different K value. The results of the classification comparison of the two methods show that the Modified K-Nearest Neighbor method, with the highest accuracy of 55.300% is superior to the K-Nearest Neighbor method with the highest accuracy of 53.300%. The two highest accuracies produced in each method were obtained using K = 5 and the proportion ratio of test data and training data is 10:90 (fold 10).

Keywords: *Classification, K-Nearest Neighbor, Modified K-Nearest Neighbor, Fold*

1. Introduction

Music genres are a commonly used way of organizing digital music [8]. Classification of genres can be easily done by listening directly to the music files manually. This manual classification has been carried out previously by experts named Tzanetakis and Cook in 2002. Currently, with the increasing number of music circulating, manual genre assignment will take a long time, so it is necessary to automatically assign genres that can help, reduce or replace roles of humans in giving genres to music [4]. Music genre classification is an important thing that has been studied for many years by the Music Information Retrieval (MIR) community since 2002 and until now, the problem of classifying music genres still continues in the Music Information Retrieval Evaluation eXchange (MIREX) which is an annual evaluation program held by an organization called the International Society for Music Information Retrieval (ISMIR).

There are several classification methods, including K-Nearest Neighbor, Support Vector Machine, Naive Bayes, and others. Of the many existing classification methods, one that is often used is K-Nearest Neighbor (K-NN). This method is able to predict the class, which in this study is called the genre, from the training data and classify the test data with K nearest neighbors [9]. In terms of accuracy, the K-NN method can still be improved because until now, improvements to the K-NN method continue to be carried out in research [8]. One of the methods resulting from the improvement of the K-NN method is the Modified K-Nearest Neighbor (MK-NN) method. The basic difference that distinguishes MK-NN from K-NN is the addition of a validation function on training data and weight voting on all test data using data validity [15]. With the validation and weighting process on the MK-NN method, it will produce better accuracy if the value of both processes is high. With these two processes, it is hoped that the Modified K-Nearest Neighbor (MK-NN) method can correct any deficiencies in the accuracy calculation process in the K-NN method [15].

In the classification of music genres, the first thing to do before classifying is to extract features from the music itself. In this study, there are 5 features extraction methods to be used for extraction which is MFCC, chroma frequencies, spectral centroid, spectral roll-off, and zero crossing rate. The five features used are from research reference which conducted to classify music and all the five features have a better result in classifying lot of music data [12].

In a previous research, study about music genre classification, has been conducted by [12] using 5 feature extraction methods (MFCC, chroma frequencies, spectral centroid, spectral roll-off, and zero crossing rate) to classify songs to 9 different genres. In this study, there are 3 classification methods used with different accuracy results namely K-Nearest Neighbor reaching 64%, Linear Support Vector Machine reaching 60%, and Poly Support Vector Machine reaching 78%.

In other studies, conducted by [10], compares two methods which is K-NN and MK-NN in identifying dental and oral diseases. This study uses 6 classes and proves that the training data is 100 and the test data is 30 and the value of K = 3, the MK-NN method can identify types of dental and oral diseases by reaching 76.66% while K-NN reaches 30%. From this study, MK-NN having higher accuracy than K-NN. This is due to the calculation of the validity value which will affect the weight voting and also the accuracy of the MK-NN itself.

Based on both studies above, authors aim to compare the K-NN method with the MK-NN method on the classification of music genres that utilize the GTZAN dataset according to the [12] as reference hoping that with the use of MK-NN method, the accuracy of K-NN method can be improved.

2. Research Methods

In this research, the audio data that will be used is GTZAN dataset which has 10 genres with 100 songs for each genre. The process of how the system will works can be seen in Figure 1.

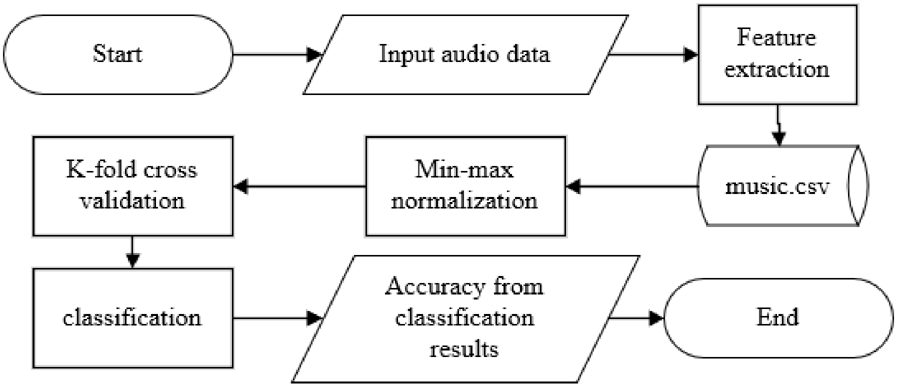


Figure 1. Flowchart of music genre classification

The process starts from feature extraction to get the features of all songs from GTZAN dataset. The features obtained will be saved into a dataset. Next step iswhatgmai to normalize the dataset using min-max normalization technique. Normalized data then will be divided equally by using K-fold cross validation technique so that each data have chances to become test data and train data. After cross validation method, data will be classified using both K-Nearest Neighbor and Modified K-Nearest

Neighbor to produce an output in the form of genre prediction which will be calculated to see how high the accuracy will be.

2.1. Research Data

Data used in this study is secondary data obtained through a website called MARSYAS which provides access to the GTZAN dataset. This GTZAN dataset has been collected and generalized by experts named Tzanetakis and Cook in 2002. The GTZAN dataset is still used by researchers in several studies, as an example of research conducted by [10] which demonstrate the fusion of visual and audio features to improve classification performance using the GTZAN dataset. Another example is research conducted by [7] using GTZAN dataset, they introduce Learning Vector Quantization (LVQ) that combined with Self Organizing Map (SOM) based on feature of entropy of wavelet coefficients which results a better accuracy than using LVQ alone.

The GTZAN dataset has a collection of 1000 songs with a duration of 30 seconds for each song. There are 10 different genres with 100 songs for each genre. All song collections are 22050Hz Mono 16-bit in *.wav format.

2.2. Feature Extraction

The goal of this step is to get a new dataset in the form of numeric values of each feature extracted from each song. There are 5 extraction features that will be used namely MFCC, chroma frequencies, spectral centroid, spectral roll-off, and zero crossing rate. Several libraries are involved for feature extraction namely scipy, librosa, and pandas.

a. Mel Frequency Cepstrum Coefficient (MFCC)

MFCC is a series of short-term power spectrum in an audio file. The MFCC models the characteristics of the human voice. The feature vector output from this extraction reaches up to 39 feature vectors [14]. In this study, 13 feature vectors were taken [12].

b. Chroma Frequencies

Chroma frequencies is one of the features that discretizes the spectrum into chromatic notes or keys and represents each note or key totaling 12. The number of feature vectors obtained is obviously will be 12 feature vectors. This feature provide a strong way to describe the similarities between one audio to another [12].

c. Spectral Centroid

This feature characterizes the signal spectrum and indicates the location of the center of gravity of the magnitude spectrum. In human perception, this is like giving the bright impression of a sound. The spectral centroid can be evaluated as a weighted average of the spectral frequencies. The higher the value of this centroid, the higher the brightness of the high frequency sound [12].

d. Spectral Roll Off

Spectral roll off can be defined as the M bin frequency below where 85% of its magnitude distribution is concentrated. In addition to the Spectral centroid, Spectral Roll Off is also a measurement of the spectral shape of the audio [12].

e. Zero Crossing Rate

Zero crossing rate is the number of times a wave crosses 0. Usually very well used for audio that has percussion instruments such as metal and rock genres [12].

2.3. Min-Max Normalization

Because different features obtained from feature extraction method has the possibility of producing a range of data that is not well distributed, then normalization is carried out first before going to the classification stage. Min-max normalization can improve data that is outliers so as to facilitate the calculation process in the classification later [1]. The order of the min-max normalization method is:

a. Get the maximum value of each feature in a data.

b. Get the minimum value of each feature in a data.

c. Apply equation (1) for each data in each existing feature.

$$norm(x) = \frac{x - minValue}{maxValue - minValue} \tag{1}$$

Where norm(x) is the x value of the i-th data on the i-th feature that has been normalized. The minValue and maxValue variables are the minimum value of the i-th feature and the largest value of the unnormalized i-th feature, and x is the unnormalized data.

2.4. Classification

Classification is done after we have a dataset either before or after performing the feature selection.

a. **K-Nearest Neighbor**

K-Nearest Neighbor works by classifying test data by studying its proximity to training data. This class of test data is obtained from the majority of classes in K data, K is the number of data with the closest distance or can also be referred to as the closest neighbor of the test data [11]. There are many distance calculations available in the K-Nearest Neighbor method, one of which is Euclidean. The purpose of the calculation is to define the distance between the two points which is the point in the training data (x) and the point in testing data (y). Calculation of euclidean distance can be done with equations (2).

$$d(xi, yi) = \sqrt{\sum_{i=0}^{n} (xi - yi)^2} \tag{2}$$

Where d is the distance between the point on the training data x and the testing data y to be classified. x, y, and i represent attribute and n is the dimension of the attribute.

b. **Modified K-Nearest Neighbor**

The Modified K-Nearest Neighbor (MK-NN) algorithm is the development of the K-Nearest Neighbor method with the addition of several processes, namely the calculation of validity values and weight calculations [13].

The MK-NN method is not much different from the K-NN which also has a distance calculation process and calculates the proximity between the test data and the training data, the amount of which depends on the K parameter. to each result from the closest distance between the test data and the training data. In the K-NN algorithm, the most class of distance calculations will be chosen as the result of the K-NN classification, but of course the classification results are not necessarily correct. Therefore, each result of the distance calculation is given a weight so that the largest weight will be selected as the classification class. In addition to calculating the weights, it is also necessary to calculate the similarity between the training data and calculate the validity value to support the weight calculation process later [3].

Each data in the training data must be validated before carrying out the next process. The validity of each data depends on each of its neighbors. The validation process is carried out for all data on the training data which then, the validity value generated from the process will be used as more information to carry out the weight calculation process later (Wafiyah et al, (2017). Calculation of validity can be seen in equation (3).

$$Validity(x) = \frac{1}{H}\sum_{i=0}^{H} S(label(x), label(Ni(x))) \tag{3}$$

Where:

H is the number of points or nearest neighbors. In this case H can be determined by the researchers themselves [5] or can be equated in value with the K value in the K-NN method [6].

label(x) is class x

label(Ni(x)) is the class from the i-th point closest to x

S is used to calculate the similarity between point x and the i-th data from the nearest neighbor of x. This S function can be defined in equation (4).

$$S(a,b) = \begin{cases} 1 \ (a = b) \\ 0 \ (a \neq b) \end{cases} \tag{4}$$

Where:

a is the class of the training data x

b is the class of the i-th training data closest to x

The weight voting stage is carried out after getting the validity value and the closest Euclidean distance from the i-th test data. Weight voting can be calculated by equation (5).

$$W_{(i)} = \ Validitas(i)\ x \frac{1}{de + a} \tag{5}$$

Where:

$W_{(i)}$ is the weight of the i-th neighbor (training data)

$Validity(i)$ is the validity value of the i-th neighbor (training data)

de is the value of the Euclidean distance from the test data to the i-th neighbor (training data)

a is the smoothing parameter [5].

Calculation of weight voting is needed in the MK-NN method to provide additional information in determining the class of test data.

2.5. K-Fold Cross Validation

K-fold cross validation is one of the techniques that can be used to divide the proportion between test data and training data. K-fold cross validation can be referred to as rotation estimation which is used to make model predictions and estimate how accurate a predictive model is when run in practice. The K-fold cross validation technique breaks the data into K parts of data sets of the same size in order that each data has the opportunity to become test data and training data. This technique can reduce bias in the data. Training and testing were carried out K-times [2].

3. Result and Discussion

3.1. Comparative Test of H Parameters Using the Modified K-Nearest Neighbor Method

The test results for the first test scenario are to determine the effect of different H values on accuracy and find the best H value for calculating the validity of the MK-NN method which can be seen in Figure 2.

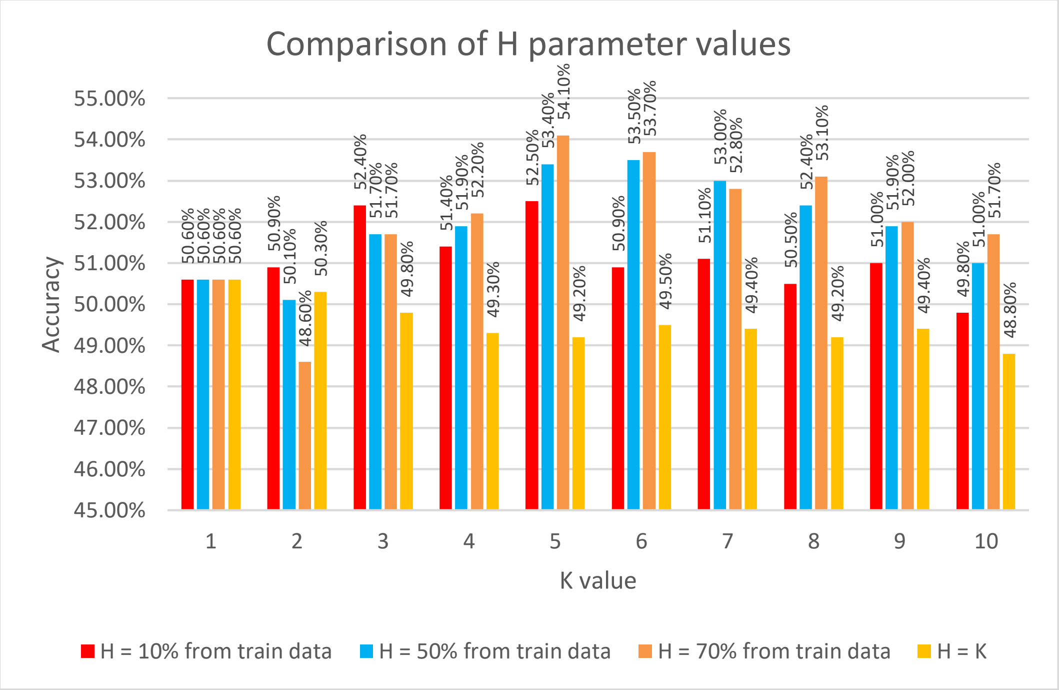


Figure 2. Bar Graph of Test Results to Determine Best H Value for MK-NN Method

Parameter H with a value of 10% from the training data got the lowest to the highest accuracy with an accuracy range of 49,800% - 52,500%, the highest accuracy of 52,500% was obtained using K = 5.

Parameter H with a value of 50% of the training data got the lowest to the highest accuracy with an accuracy range of 50.100% - 53.500%, the highest accuracy of 53.500% was obtained using K = 6.

Parameter H with a value of 70% from the training data got the lowest to the highest accuracy with an accuracy range of 48.600% - 54.100%, the highest accuracy of 54.100% was obtained using K = 5.

The H parameter with the same value as the K value gets the lowest to the highest accuracy with an accuracy range of 48.800% - 50.600%, the highest accuracy of 50.600% is obtained by using K=1 When viewed from the highest average, the best H parameter is obtained with a value of 70% from the training data with K = 5, namely with an accuracy of 54.100%, so that this H value will then be used to calculate the validity of the second and third test scenarios.

3.2. Comparison Test of K Parameters with a Proportion Ratio of 20:80 (Fold 5) Using the K-Nearest Neighbor Method and the Modified K-Nearest Neighbor Method

The test results for the second test scenario is to see the effect of different K parameters using the proportion of test data and training data 20:80 (fold 5) can be seen in the bar graph in Figure 3.

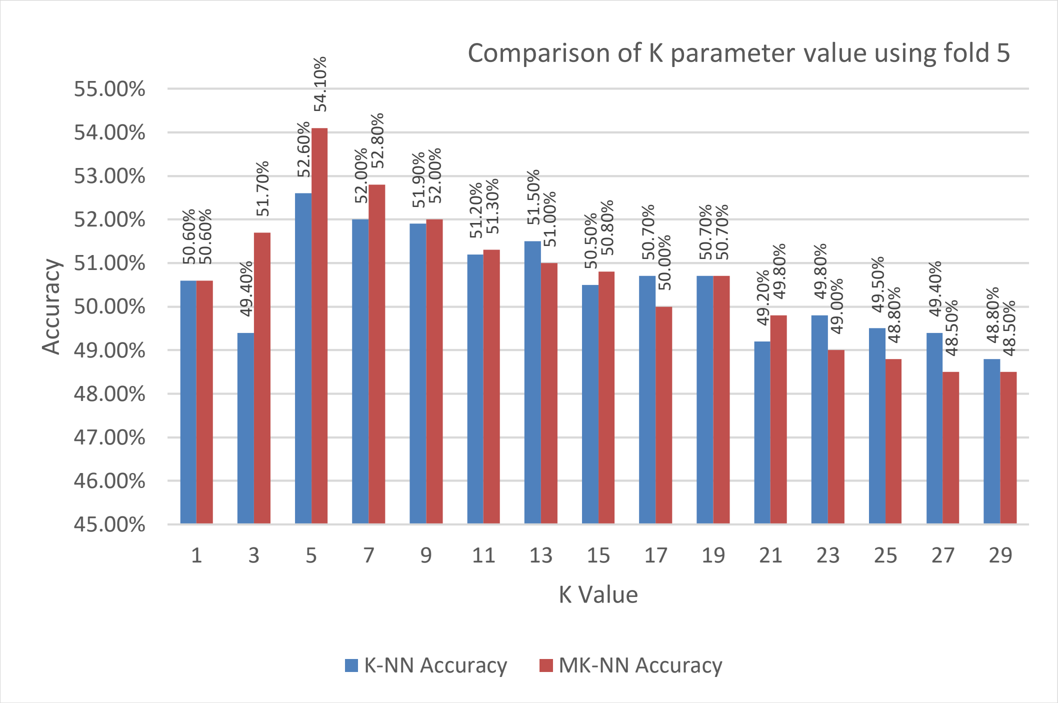


Figure 3. Bar Graph of Test Result Using Fold 5

Results shown on Figure 3 indicate that the K-NN method gets accuracy from the lowest to the highest with an accuracy range of 48.800% - 52.600% and the MK-NN method gets accuracy from the lowest to the highest with an accuracy range of 48.500% - 54.100%

The highest accuracy for the K-NN method with fold 5 was obtained at the value of K = 5 with an accuracy of 52.600% and the lowest accuracy was obtained at the values of K = 29 with an accuracy of 48.800%.

The highest accuracy for the MK-NN method with fold 5 was obtained at the value of K = 5 with an accuracy of 54.100% and the lowest accuracy was obtained at the value of K = 27 and 29 with an accuracy of 48.500%.

The bar graph in Figure 3 shows that the accuracy obtained at the value of K = {1, 3, 5, 7, 9, 11, 13, 15, 17, 19} still does not show a consistent decrease and the highest peak of accuracy is at K = 5 for the K-NN and MK-NN methods. Accuracy begins to show a gradual decrease at the value of K = 21 and so on. The total average accuracy of the K-NN method is 50.520%, while the total average accuracy of the MK-NN method is 50.640% which proves that the MK-NN method outperforms the K-NN method.

3.3. Comparison Test of K Parameters with a Proportion Ratio of 10:90 (Fold 10) Using the K-Nearest Neighbor Method and the Modified K-Nearest Neighbor Method

The test results for the second test scenario is to see the effect of different K parameters using the proportion of test data and training data 10:90 (fold 10) can be seen in the bar graph in Figure 4.

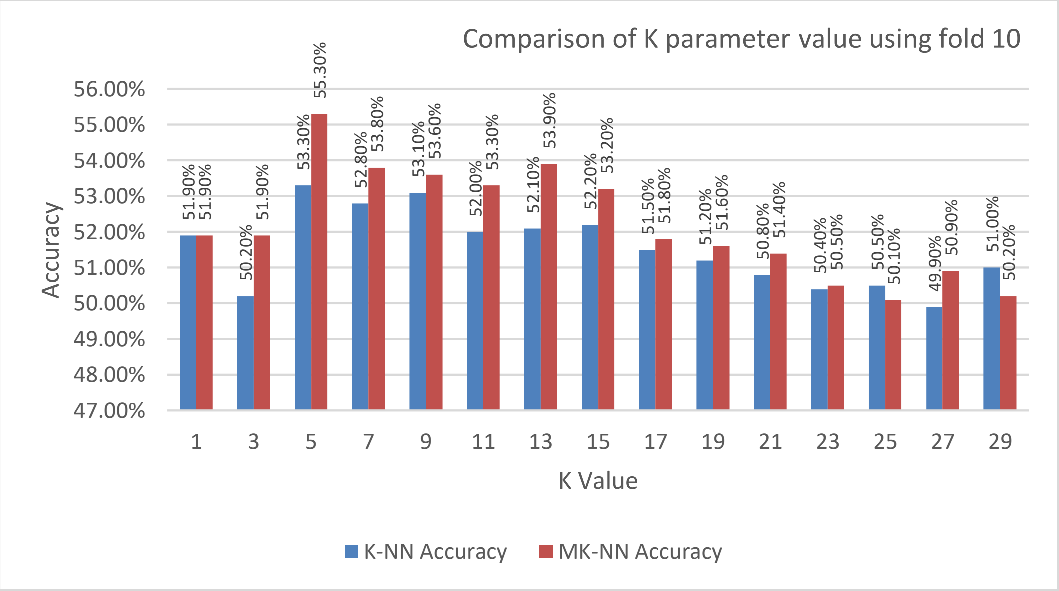


Figure 4. Bar Graph of Test Result Using Fold 10

Results shown on Figure 4 indicate that the K-NN method gets accuracy from the lowest to the highest with an accuracy range of 49.900% - 53.300% and the MK-NN method gets accuracy from the lowest to the highest with an accuracy range of 50.100% - 55.300%

The highest accuracy for the K-NN method with fold 10 was obtained at the value of K = 5 with an accuracy of 53.300% and the lowest accuracy was obtained at the values of K = 27 with an accuracy of 49.900%.

The highest accuracy for the MK-NN method with fold 10 was obtained at the value of K = 5 with an accuracy of 55.300% and the lowest accuracy was obtained at the value of K = 25 with an accuracy of 50.100%.

The bar graph in Figure 4 shows that the accuracy obtained at the value of K = {1, 3, 5, 7, 9, 11, 13, 15} still does not show a consistent decrease and the highest peak of accuracy is at K = 5 for the K-NN and MK-NN methods. Accuracy begins to show a gradual decrease at the value of K = 17 and so on. The total average accuracy of the K-NN method is 51.527%, while the total average accuracy of the MK-NN method is 52.227% which proves that the MK-NN method outperforms the K-NN method, same as previous scenario using fold 5.

3.4. Test Result Analysis

In Figure 2, several values for the H parameter are tested to find the best value for calculating the validity of the Modified K-Nearest Neighbor method and from the results of the first scenario testing that tests this H parameter, the best H value is sought by looking at the highest accuracy generated from several K values that have been determined for each H value tested. The highest average was obtained with a value of H = 70% of the training data using K = 5, this indicates that for this study, with a large number of training data (800 training data), a large H value tends to provide a better validity value. in the Modified K-Nearest Neighbor method which is able to increase the accuracy obtained from the method. The best H value is then used for testing the next test scenario.

In Figure 3, a test was conducted with several K values with a ratio of the proportion of test data and training data of 20:80 and in Figure 4, a test was carried out with several K values with a ratio of the proportion of test data and training data of 10:90. From the two tests, the overall accuracy results are superior to the Modified K-Nearest Neighbor method which obtains the highest accuracy, namely 55.300% with a K = 5 value and using a data proportion of 10:90 while the highest accuracy by the K-

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Nearest Neighbor method is 53.300% with value of K = 5 and by using the proportion of data 10:90. Both tests show that the proportion of test data and training data is able to affect the classification results obtained from the K-Nearest Neighbor and Modified K-Nearest Neighbor methods. The K value can also affect the classification results where the MK-NN method is able to maintain its superiority compared to the K-NN method when given a large K value even though both methods continue to experience a decrease in accuracy where this can be caused by weight calculations that strengthen the MK-NN method in selecting the prediction class from the tested data.

4. Conclusion

This study has succeeded in classifying using the K-Nearest Neighbor and Modified K-Nearest Neighbor methods. Several conclusions can be drawn based on the results of research that has been carried out. The first test scenario is to know the effect of the value on the H parameter for validation calculations on the Modified K-Nearest Neighbor method show that the H value can affect the accuracy of the method. The best H value was 70% of the training data, the use of this H value was able to get the highest accuracy of 54.100% with a K = 5 value and the ratio of the proportion of test data and training data was 20:80 in the Modified K-Nearest Neighbor method. The second test is to know the effect of the value on the K parameter and the proportion of different test data and training data on the K-Nearest Neighbor and Modified K-Nearest Neighbor methods indicate that the K parameter and the proportion of data are able to affect the accuracy of the two methods. The highest accuracy obtained in the Modified K-Nearest Neighbor method is 55.300% using the value of K = 5 and the ratio of the proportion of test data and training data is 10:90, while the highest accuracy obtained by the K-Nearest Neighbor method is 53.300% using K value and the ratio of the proportion of test data and training data are the same. The third scenario test is the comparison of the results in classification test between the K-Nearest and Modified K-Nearest Neighbor methods which show that overall, the Modified K-Nearest Neighbor method is able to outperform the K-Nearest Neighbor method in terms of accuracy and both methods get the highest accuracy of each, namely 53,300%. for K-Nearest Neighbor and 55.300% for Modified K-Nearest Neighbor using the value of K = 5 and the ratio of the proportion of test data and training data 10:90.

References

[1] D. A. Nasution, H. H. Khotimah, N. Chamidah, “Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma K-NN” *Computer Engineering System and Science*, Vol. 4, No. 1, p. 78-82, 2019.

[2] F. Tempola, M. Muhammad, A. Khairan, “*Perbandingan Klasifikasi Antara KNN dan Naïve Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation*” *Jurnal Teknologi Informasi dan Ilmu Komputer*, Vol. 5 Issue.5, 2018.

[3] F. Wafiyah, N. Hidayat, R.S. Perdana, “*Implementasi Algoritma Modified KNearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam*” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2017.

[4] Gst. A. V. M. Giri, “*Klasifikasi dan Retrieval Musik Berdasarkan Genre (Sebuah Studi Pustaka)*” *Jurnal Ilmiah Ilmu Komputer Universitas Udayana*, 2017.

[5] H. Parvin, H. Alizadeh, B. Minati, “*A Modification on K-Nearest Neighbor Classifier*” *Global Journal of Computer Science and Technology,* Vol. 10, Issue 14, 2010.

[6] I. Agustin, Y.N. Nasution, Wasono, “*Klasifikasi Batubara Berdasarkan Jenis Kalori Dengan Menggunakan Algoritma Modified K-Nearest Neighbor (Studi Kasus: PT.Pancaran Surya Abadi)*” *Jurnal EKSPONENSIAL,* Vol. 10, Issue.1, 2019.

[7] L. A. A. R. Putri, S. Hartati, *"Klasifikasi Genre Musik Menggunakan Learning Vector Quantization dan Self Organizing Map" Jurnal Ilmiah ILMU KOMPUTER Universitas Udayana,* Vol. 9, No. 1, p. 14-22, 2016.

[8] L. Nanni, Y.M.G. Costa, A. Lumini, M.Y. Kim, S.R. Baek, *"*Combining visual acoustic features for music genre classification*"* Expert System With Applications, p. 1-10, 2016.

[9] M. Holeňa, P. Pulc, M. Kopp, "*Classification Methods for Internet Applications*" *Poland: Springer International Publishing*, 2020.

[10] M.R. Ravi, Indriati, S. Adinugroho, "*Implementasi Algoritma Modified KNearest Neighbor (MKNN) Untuk Mengidentifikasi Penyakit Gigi dan Mulut*" *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2019.

[11] N. Fetra, M. Irsyad, "*Aplikasi Pencarian Chord dalam Membantu Penciptaan Lagu Menggunakan Algoritma Fast Fourier Transform (FFT) dan Metode Klasifikasi K-Nearest Neighbor (KNN)*" *Jurnal CoreIT,* Vol. 1, Issue.2, 2015.

[12] N. M. Patil, M. U. Nemade, "Music Genre Classification Using MFCC, K-NN, and SVM Classifier" *Computer Engineering in Research Trends*, Vol. 4, Issue 2, p. 43-47, 2017.

[13] N. Wahyudi, S. Wahyuningsih, F.D.T. Amijaya, "*Optimasi Klasifikasi Batubara Berdasarkan Jenis Kalori dengan menggunakan Genetic Modified K-Nearest Neighbor (GMK-NN) (Studi Kasus: PT Jasa Mutu Mineral Indonesia Samarinda, Kalimantan Timur*" *Jurnal EKSPONENSIAL*, 2019.

[14] S.A. Majeed, H. Husain, S.A. Samad, T.F. Idbeaa, "*Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: A Comparison Study. Journal of Theoretical and Applied Information Technology*" *Journal of Theoretical and Applied Information Technology*, 2015.

[15] T. H. Simanjuntak, W. F. Mahmudy, Sutrisno, "Implementasi *Modified K-Nearest Neighbor* Dengan Otomatisasi Nilai K Pada Pengklasifikasian Penyakit Tanaman Kedelai" *Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol. 1, No. 2, p. 75-79, 2017.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Analisis Forensik Digital pada Aplikasi Twitter di Android sebagai Bukti Digital dalam Penanganan Kasus Prostitusi Online

Sang Putu Febri Wira Pratama[a1], I Gusti Ngurah Anom Cahyadi Putra[a2], Muhammad Akbar Hamid[a3], Calvin Christian[a4], I Ketut Kusuma Merdana[a5]

[a]Informatics Department, Udayana University
Jalan Raya Kampus Unud, Jimbaran, Bali, 80361, Indonesia
[1]febriwiraprtma@gmail.com
[2]anom.cp@unud.ac.id
[3]m_akbarhamid@student.unud.ac.id
[4]calvinchristian15k1@gmail.com
[5]ketutkusuma0910@gmail.com

Abstract

At this time the use of social media from time to time has experienced rapid development, one of which is the social media twitter. Twitter social media has many benefits such as making tweets about daily activities. However, Twitter social media has a negative side in its use, one of which is online prostitution. Prostitution is an act of cyber crime that violates the rules and norms that exist in society. Therefore, to overcome these cyber crimes, the necessary action is to review online prostitution on Twitter social media. In this study, a digital forensic analysis was conducted on Twitter social media on smartphones related to acts of prostitution using the National Institute of Justice (NIJ). Based on the research conducted, digital evidence is obtained that can be accounted for by the perpetrators.

Keywords: *Digital Forensics, Smartphone, Twitter, Prostitution, National Institute of Justice Method*

1. Pendahuluan

1.1. Latar Belakang

Pada saat ini penggunaan dari media sosial dari waktu ke waktu mengalami perkembangan yang pesat. Media sosial merupakan kebutuhan tersier yang kebutuhannya harus dipenuhi seperti kebutuhan primer, dan dapat dinikmati sesuai dengan kebutuhan pengguna. Berdasarkan penelitian yang ditulis oleh organisasi We Are Social and Hootsuite pada tahun 2020 mengatakan bahwa pengguna jejaring sosial di seluruh dunia adalah 3.534 miliar orang atau 46% dari total populasi dunia dan sekitar 3.463 miliar orang mengakses media sosial melalui *smartphone* atau 45% dari populasi dunia [1].

Media sosial yang paling banyak dalam penggunaannya salah satunya adalah twitter. Media sosial twitter diciptakan tahun 2006 di San Fransisco, Amerika Serikat. Berdasarkan penelitian yang ditulis oleh KEMENKOMINFO atau Kementrian Komunikasi dan Informatika menyatakan pada saat ini Indonesia berada pada peringkat 5 dari seluruh dunia pada pengguna twitter. Berdasarkan data yang dimiliki oleh PT Bakrie Telecom, Indonesia mempunyai setidaknya 19,5 juta pengguna twitter dari keseluruhan pengguna twitter dunia yang berkisar 500 juta [2].

Media sosial twitter memiliki banyak manfaat, salah satunya adalah dengan membuat konten cerita, sebagai tempat mencurahkan isi hati, atau sebagai tempat berbagi konten positif. Akan tetapi di sisi lainnya, media sosial twitter memiliki sisi negatif, contohnya adalah banyak dan mudahnya tersebar konten pornografi, dan prostitusi online. Prostitusi merupakan tindakan kejahatan *cyber* yang menyalahi aturan dan norma yang ada di masyarakat. Maka dari itu, untuk mengatasi tindakan kejahatan *cyber* tersebut, diperlukannya tindakan forensik untuk mengatasi prostitusi online yang berada di media sosial twitter.

1.2. Rumusan Masalah

Rumusan masalah pada penelitian yang dilakukan adalah bagaimana hasil akhir dari akuisisi yang dilakukan menggunakan analisa forensik pada media sosial twitter sebagai bukti digital untuk kasus-kasus tindak kejahatan prostitusi online.

1.3. Tinjauan Pustaka

a. Twitter

Twitter adalah media sosial berbentuk *microblogging* dikarenakan dalam membuat *posting* terdapat batasan 140 karakter. *Posting*-an yang berada di twitter disebut *tweet* dan terdapat pada kamus Oxford English Dictionary (OED) [3]. Aplikasi mobile twitter yang terdapat pada smartphone android memiliki tempat penyimpanan data yang terdapat pada folder com.twitter.android atau jika mengakses secara lengkap berada pada root/data/data/com.twitter.android. Untuk mendapatkan folder tersebut harus menggunakan folder *root*, yang mana untuk mengaksesnya diperlukan perangkat android yang telah di *rooting* [4].

b. Prostitusi Online

Prostitusi atau biasa disebut sebagai pelacuran yang ditulis dalam Kamus Besar Bahasa Indonesia (KBBI) mengatakan bahwa kata prostitusi atau pelacuran berasal dari kata lacur yang memiliki arti malang, sial, celaka, buruk laku, atau gagal. Sedangkan untuk pelacur memiliki arti perempuan atau wanita tuna susila, sundal, atau melacur [5].

Prostitusi dapat dilakukan dimana dan kapanpun di seluruh dunia. Di negara lain, prostitusi dilakukan secara sembunyi-sembunyi atau gelap pada rumah pelacuran atau di rumah-rumah [6]. Prostitusi online merupakan penyakit yang terdapat di masyarakat dimana perempuan menjual diri mereka, dan melakukan tindakan seksual dan media sosial online yang mana sebagai perantara untuk membantu mempromosikan.

c. Forensik Digital

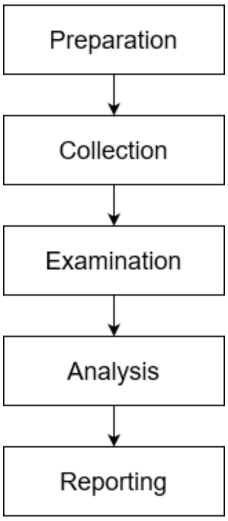
Forensik digital adalah suatu langkah dalam memeriksa media digital atau perangkat yang berhubungannya seperti smartphone, harddisk, folder, yang menggunakan metode untuk mengolahnya. Dapat dikatakan forensik digital adalah suatu langkah untuk menjaga, mengumpulkan, menyimpan, menganalisa dan menyajikan suatu barang bukti yang berhubungan dengan obyek digital. Forensik digital adalah suatu langkah untuk mendapatkan dan menganalisa informasi yang berbentuk digital untuk digunakan dalam pengajuan barang bukti di pengadilan [7].

d. Bukti Digital

Bukti digital merupakan informasi yang dapat dijadikan barang bukti yang sah di pengadilan. Untuk barang bukti digital yang berhubungan dengan *smartphone* atau *mobile* dapat ditemukan di *history chat*, log, audio, foto dan lain-lain. Bukti digital umumnya dikaitkan dengan kejahatan yang terjadi di dunia maya seperti kejahatan yang menggunakan perantara media sosial, sehingga untuk mengadili kasus kejahatan tersebut sangat penting untuk menggunakan bukti digital sebagai bukti [8]. Bukti digital sangat lemah untuk mempertahankan keasliannya jika tidak ditangani dengan baik. Keaslian atau tidak yang terdapat pada bukti digital dapat digunakan sebagai bukti atau kesimpulan yang berguna atau tidak berguna [9].

2. Metode Penelitian

Dalam penelitian yang dilakukan menggunakan metode *National Institute of Justice* (NIJ) yang berfungsi untuk melakukan tahapan atau alur sehingga dapat dijadikan acuan untuk pemecahan masalah. Tahapan metode NIJ dijelaskan sebagai berikut:



Gambar 1. Metode *National Institute of Justice (NIJ)*

a. *Preparation*

Tahap *preparation* atau persiapan merupakan proses dalam pemilahan barang-barang yang digunakan sebagai barak bukti tindak kriminalitas. Barang yang digunakan dapat berbentuk barang digital atau perangkat keras dan barang bukti tersebut dapat digunakan dalam proses penyidikan.

b. *Collection*

Tahap *collection* atau pengumpulan adalah tahapan yang digunakan untuk mengumpulkan data yang dapat digunakan dalam proses penyidikan. Pada proses ini terdapat pengambilan data-data yang terdapat pada sumber pada barang bukti dan menjaga keaslian serta keutuhan barang bukti tersebut dari perubahan.

c. *Examination*

Tahap examination atau pemeriksaan adalah tahapan pemeriksaan data yang telah dikumpulkan secara forensik, dan memastikan data yang digunakan merupakan data yang asli.

d. *Analysis*

Tahap *analysis* atau analisis merupakan proses untuk mendapatkan data atau file digital yang dapat digunakan sebagai bukti dari proses pemeriksaan, dan selanjutnya data yang didapatkan akan dianalisis secara detail dan menyeluruh dengan metode yang sah secara hukum dan teknik sebagai bentuk keadilan dalam pengungkapan barang bukti digital. Hasil dari tahap analisis digunakan sebagai bukti digital yang dapat dipertanggungjawabkan secara hukum dan ilmiah.

e. *Reporting*

Tahap *reporting* atau pelaporan adalah langkah untuk pelaporan hasil dari tahapan analisis yang merupakan gambaran dari tindakan yang dilakukan, peralatan yang digunakan dalam pengungkapan barang bukti, dan metode yang digunakan.

3. Hasil dan Pembahasan

Dalam penelitian ini menggunakan simulasi kejadian kasus prostitusi online yang terjadi pada aplikasi mobile twitter. Pada contoh kasus dimana pelaku atau mucikari diamankan oleh polisi yang menyamar sebagai pelanggan, didapatkan barang bukti dari pelaku berupa smartphone dengan merk Xiaomi Redmi Note 10 Pro yang didalamnya terdapat aplikasi Twitter sebagai sarana melakukan promosi prostitusi dan transaksi. Pada aplikasi Twitter, pelaku memiliki akun dengan ID yaitu "desipuspita3009". Sebagai tindakan lebih lanjut, pihak berwajib menyita smartphone milik PSK tersebut untuk penyelidikan. Dalam penyidikan, penyidik menggunakan metode NIJ yang mempunyai lima proses dasar dalam Forensik, yaitu Preparation, Collection, Examination, Analysis dan Reporting.

3.1. Preparation

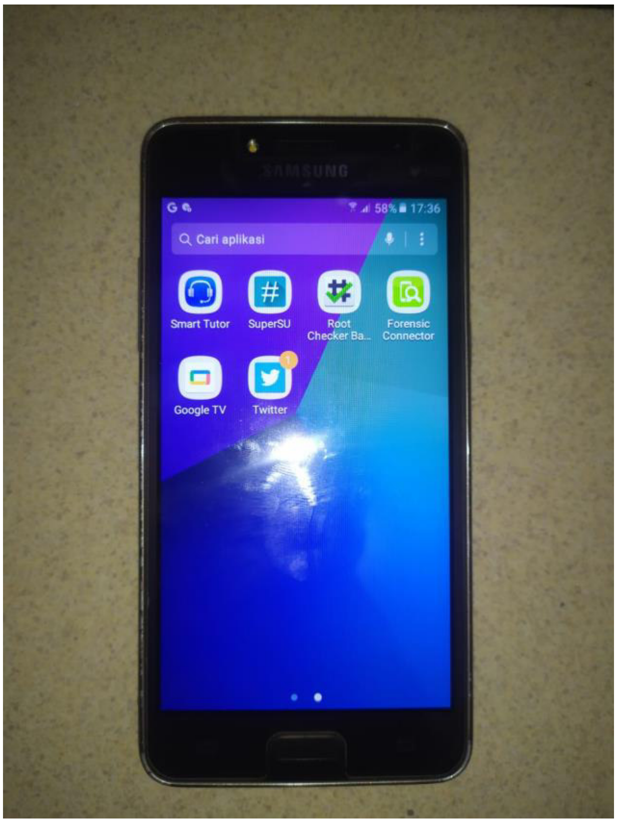
Dalam proses *preparation* atau persiapan berfungsi untuk menyiapkan alat-alat yang digunakan sebagai barang bukti dan proses penyidikan. Alat dan bukti yang digunakan terdapat pada tabel 1.

Tabel 1. Alat dan bukti

No	Jenis Perangkat	Alat dan Bukti	Spesifikasi
1	Hardware	Laptop	Lonovo Z40-75/VGA AMD A10-7300 1,9GHz/RAM 4GB
2	Hardware	Smartphone	Samsung J2 Prime
3	Software	Mobile Forensic Express	Version 7.4.1.21502 (64-bit)
4	Software	SysTools SQLite Viewer	Versi 1.2

3.2. Collection

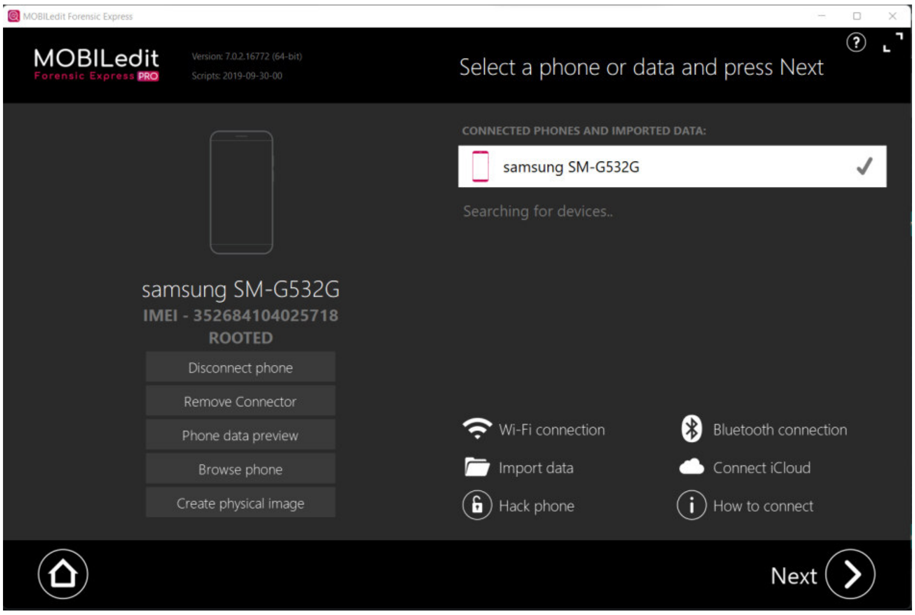
Dalam proses ini, penyidik mengumpulkan data fisik dan dokumentasi, serta mengumpulkan data yang terdapat pada smartphone terduga.



Gambar 2. Barang Bukti *Smartphone*

Pada gambar 2, merupakan dokumentasi barang bukti fisik dari alat komunikasi yang digunakan oleh terduga untuk melakukan tindakan prostitusi online berupa *smartphone* dengan merk Samsung J2 Prime. Barang bukti smartphone menggunakan sistem operasi android dengan versi 6.0 yang mana telah terpasang aplikasi media sosial twitter mobile. Selanjutnya penyidik akan mengambil data pada *smartphone* dengan cara mengkloning, guna untuk menghindari perubahan data atau penghapusan data yang nantinya akan menjadi barang bukti digital.

3.3. Examination



MOBILedit Forensic Express

MOBILedit Forensic Express PRO

Version: 7.0.2.16772 (64-bit)
Scripts: 2019-09-30-00

Select a phone or data and press Next

CONNECTED PHONES AND IMPORTED DATA:

samsung SM-G532G ✓

Searching for devices..

samsung SM-G532G
IMEI - 352684104025718
ROOTED

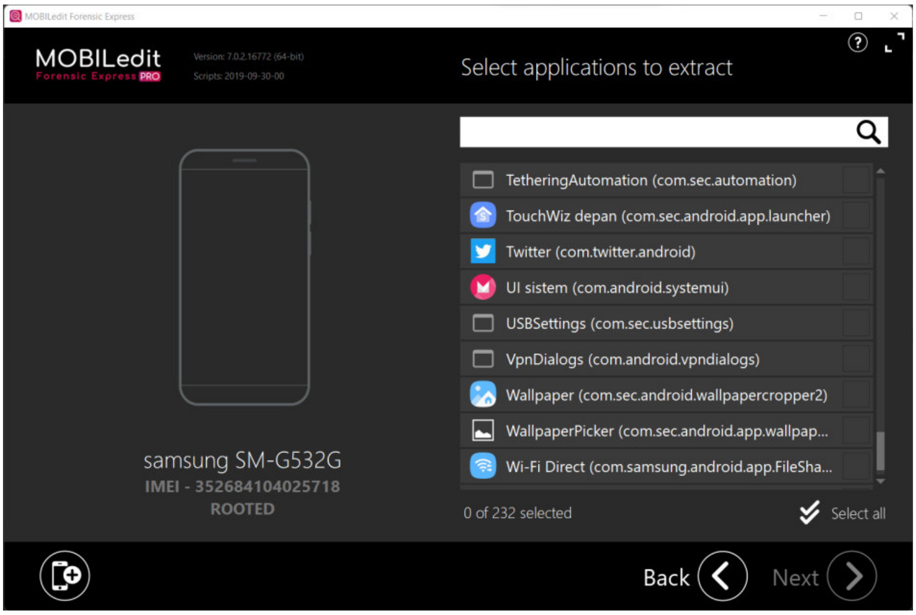
Disconnect phone
Remove Connector
Phone data preview
Browse phone
Create physical image

Wi-Fi connection
Bluetooth connection
Import data
Connect iCloud
Hack phone
How to connect

Next

Gambar 3. Informasi IMEI dan Status Root pada Samsung J2 Prime

Dalam proses ini penyidik melakukan pengecekan data pada smartphone menggunakan bantuan perangkat lunak *Mobiledit Forensic Express* yang sudah terpasang di laptop. Jika terhubung dengan smartphone pelaku, maka *Mobileedit Forensic Express* akan menampilkan informasi nomor IMEI dan Status Root dari smartphone tersebut, seperti gambar 3.



MOBILedit Forensic Express

MOBILedit Forensic Express PRO

Version: 7.0.2.16772 (64-bit)
Scripts: 2019-09-30-00

Select applications to extract

□ TetheringAutomation (com.sec.automation)
TouchWiz depan (com.sec.android.app.launcher)
Twitter (com.twitter.android)
UI sistem (com.android.systemui)
□ USBSettings (com.sec.usbsettings)
□ VpnDialogs (com.android.vpndialogs)
Wallpaper (com.sec.android.wallpapercropper2)
WallpaperPicker (com.sec.android.app.wallpap...
Wi-Fi Direct (com.samsung.android.app.FileSha...

0 of 232 selected

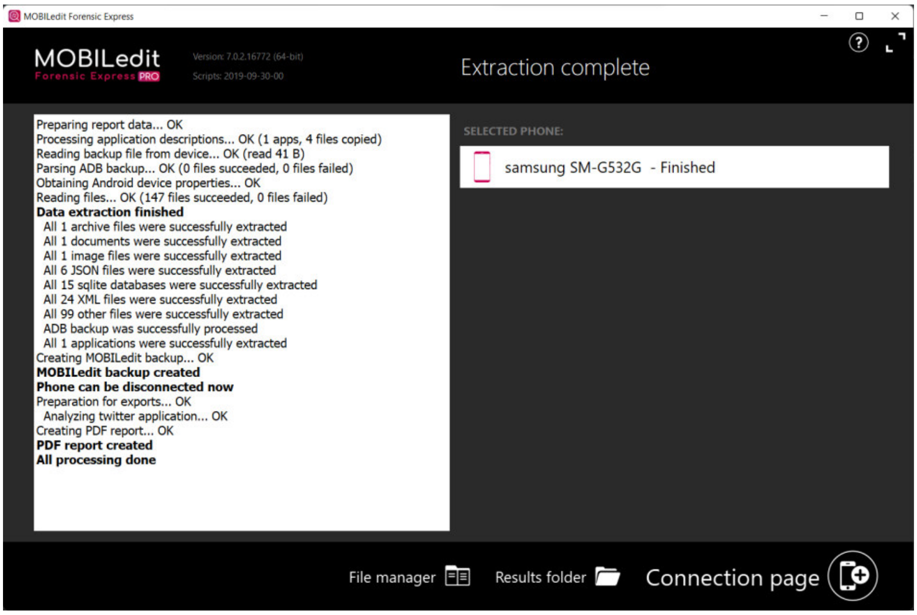
✓ Select all

samsung SM-G532G
IMEI - 352684104025718
ROOTED

Back
Next

Gambar 4. Akuisisi Data dari Aplikasi Twitter

Selanjutnya proses akuisisi data dari aplikasi Twitter yang terdapat di *smartphone* pelaku, yaitu Samsung J2 Prime yang ditunjukkan pada Gambar 4. Kemudian *extract* data dari aplikasi Twitter sehingga barang bukti dari *smartphone* tersebut didapatkan (Lihat Gambar 5).



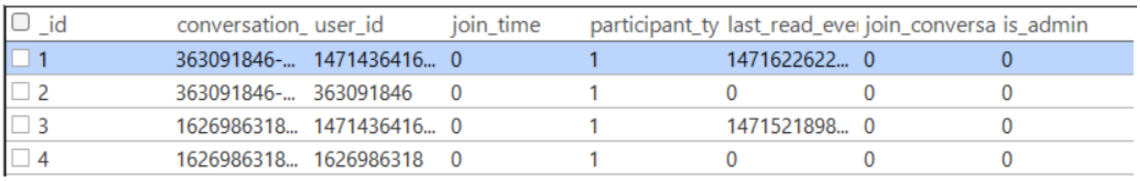
```
Preparing report data... OK
Processing application descriptions... OK (1 apps, 4 files copied)
Reading backup file from device... OK (read 41 B)
Parsing ADB backup... OK (0 files succeeded, 0 files failed)
Obtaining Android device properties... OK
Reading files... OK (147 files succeeded, 0 files failed)
Data extraction finished
  All 1 archive files were successfully extracted
  All 1 documents were successfully extracted
  All 1 image files were successfully extracted
  All 6 JSON files were successfully extracted
  All 15 sqlite databases were successfully extracted
  All 24 XML files were successfully extracted
  All 99 other files were successfully extracted
  ADB backup was successfully processed
  All 1 applications were successfully extracted
Creating MOBILedit backup... OK
MOBILedit backup created
Phone can be disconnected now
Preparation for exports... OK
  Analyzing twitter application... OK
Creating PDF report... OK
PDF report created
All processing done
```

Gambar 5. Hasil dari Proses Ekstrak

Dari hasil analisa ditemukan beberapa file berupa database dengan format SQLite. Untuk mengetahui isi file database, diperlukan software *SysTools SQLite Viewer* untuk analisis lebih lanjut.

3.4. Analysis

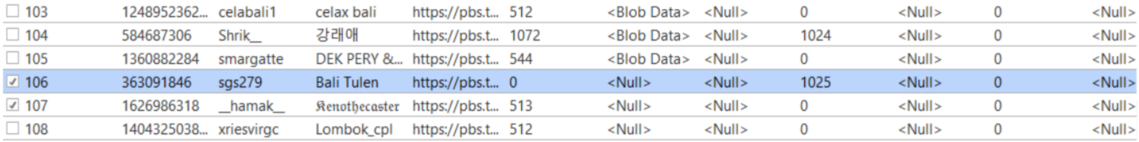
Berdasarkan analisa isi dari data yang telah dibangkitkan dari aplikasi Twitter ditemukan bahwa data penting yang dapat digunakan untuk mendukung penyelidikan adalah database yang disebut "1471436416148119552-61". Berdasarkan hasil investigasi yang terdapat pada *tools* Systools SQLite Viewer, penyidik ingin mendapatkan percakapan yang terjadi diantara pelaku dan pelanggan. Jadi penyidik melakukan pengecekan terhadap tabel conversation_participants.



□ _id	conversation_	user_id	join_time	participant_ty	last_read_evei	join_conversa	is_admin
□ 1	363091846-...	1471436416...	0	1	1471622622...	0	0
□ 2	363091846-...	363091846	0	1	0	0	0
□ 3	1626986318...	1471436416...	0	1	1471521898...	0	0
□ 4	1626986318...	1626986318	0	1	0	0	0

Gambar 6. Kolom dari Tabel Conversation_Participants

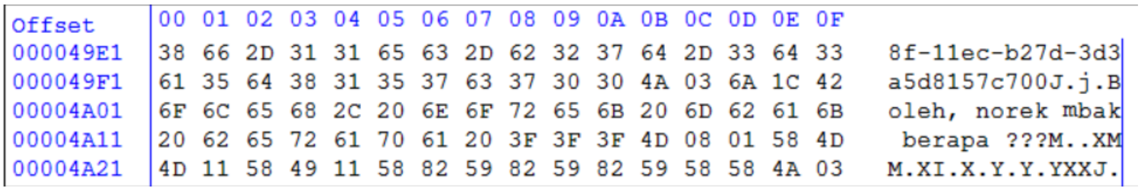
Kemudian penyidik ingin mengetahui *username* dari pemilik *user_id* yang tercatat pada tabel *conversation_entries* sehingga penyidik melakukan pengecekan terhadap tabel *user*.



□ 103	1248952362...	celabali1	celax bali	https://pbs.t...	512	<Blob Data>	<Null>	0	<Null>	0	<Null>
□ 104	584687306	Shrik_	강래애	https://pbs.t...	1072	<Blob Data>	<Null>	1024	<Null>	0	<Null>
□ 105	1360882284	smargatte	DEK PERY &...	https://pbs.t...	544	<Blob Data>	<Null>	0	<Null>	0	<Null>
☑ 106	363091846	sgs279	Bali Tulen	https://pbs.t...	0	<Null>	<Null>	1025	<Null>	0	<Null>
☑ 107	1626986318	_hamak_	𝔎𝔢𝔫𝔬𝔱𝔥𝔢𝔠𝔞𝔰𝔱𝔢𝔯	https://pbs.t...	513	<Null>	<Null>	0	<Null>	0	<Null>
□ 108	1404325038...	xriesvirgc	Lombok_cpl	https://pbs.t...	512	<Null>	<Null>	0	<Null>	0	<Null>

Gambar 7. Kolom dari Tabel User

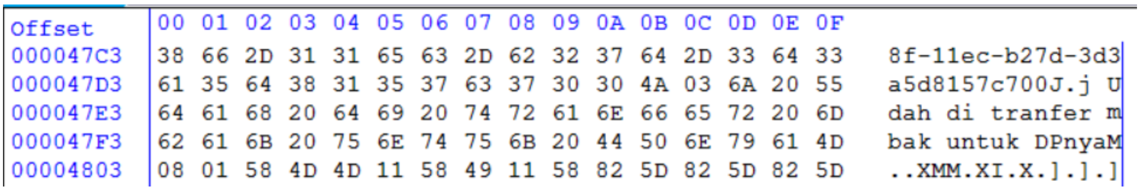
Diketahui bahwa pengguna dari *user_id* “1626986318” dan “363091846” masing-masing dengan *username* “_hamak_” dan “sgs279”. Lebih lanjut, penyidik ingin mengetahui isi percakapan dari pengguna tersebut maka penyidik melakukan pengecekan terhadap tabel *conversation_entries*. Penyidik kemudian mendapatkan bukti bahwa pengguna dengan *username* “_hamak_” sedang melakukan transaksi gelap dengan pelaku yang ditunjukkan pada Gambar 8.



```
Offset    00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
000049E1  38 66 2D 31 31 65 63 2D 62 32 37 64 2D 33 64 33   8f-11ec-b27d-3d3
000049F1  61 35 64 38 31 35 37 63 37 30 30 4A 03 6A 1C 42   a5d8157c700J.j.B
00004A01  6F 6C 65 68 2C 20 6E 6F 72 65 6B 20 6D 62 61 6B   oleh, norek mbak
00004A11  20 62 65 72 61 70 61 20 3F 3F 3F 4D 08 01 58 4D    berapa ???M..XM
00004A21  4D 11 58 49 11 58 82 59 82 59 82 59 58 58 4A 03   M.XI.X.Y.Y.YXXJ.
```

Gambar 8. Percakapan antara Pelaku dengan Pengguna (_hamak_)

Pada gambar di atas, terjadi percakapan antara pelaku dengan salah satu pelanggan yang saling melakukan transaksi Prostitusi Online. Bukti lainnya juga ditunjukkan pada Gambar 9. Selanjutnya hasil analisis tersebut dapat dijadikan sebagai bukti digital kasus Prostitusi Online dengan menggunakan aplikasi media sosial Twitter.



```
Offset    00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
000047C3  38 66 2D 31 31 65 63 2D 62 32 37 64 2D 33 64 33   8f-11ec-b27d-3d3
000047D3  61 35 64 38 31 35 37 63 37 30 30 4A 03 6A 20 55   a5d8157c700J.j U
000047E3  64 61 68 20 64 69 20 74 72 61 6E 66 65 72 20 6D   dah di tranfer m
000047F3  62 61 6B 20 75 6E 74 75 6B 20 44 50 6E 79 61 4D   bak untuk DPnyaM
00004803  08 01 58 4D 4D 11 58 49 11 58 82 5D 82 5D 82 5D   ..XMM.XI.X.].].]
```

Gambar 9. Bukti Lain Terkait Prostitusi Online

3.5. Reporting

Metode yang diterapkan dalam penyidikan ini menggunakan metode NIJ yang memiliki 5 tahapan dasar yaitu persiapan alat yang digunakan meliputi laptop, *smartphone* pelaku, aplikasi *Mobiledit Forensic Express* dan *Systool SQLite Viewer*, pengumpulan barang bukti fisik yaitu *smartphone* pelaku kemudian data tersebut digandakan agar keutuhan data tetap terjaga, pemeriksaan terhadap data yang terdapat pada smartphone pelaku, yang kemudian akan dilakukan analisis lebih dalam, kemudian akan dilakukan analisis yang lebih mendalam terhadap *database* “1471436416148119552-61”. Dari hasil investigasi database memiliki 40 tabel. Penyidik melakukan penyidikan mulai dari tabel *conversatio_participants*, kemudian dilanjutkan tabel *user* dan terakhir pada tabel *conversation_entries* penyidik menemukan bukti adanya tindak Prostitusi Online.

4. Kesimpulan

Penelitian terkait digital forensik terkait Prostitusi Online pada aplikasi Twitter ini memakai metode NIJ dengan tahapan dari *preparation*, *colletion*, *examination, analysis*, dan *reporting*. Tujuan dari penelitian ini adalah untuk membantu penyidik menyelesaikan masalah prostitusi online pada aplikasi media Twitter dengan menggunakan perangkat lunak *Mobiledit Forensic Express* dan *Systools SQLite Viewer*. Hasil dari penelitian ini dimana penyidik menemukan percakapan terkait tindak prostitusi online pada tabel *conversation_entries* dalam *database* “1471436416148119552-61”. Selanjutnya, hasil analisis tersebut dapat digunakan sebagai barang bukti digital dan dapat dipertanggungjawabkan oleh pelaku.

Referensi

[1] S. Kemp, “Essential Insights into How People around the World Use the Internet, Mobile Device, Social Media, and Ecommerce.” Hootsuite, 2020.

[2] Kominfo, “Pengguna Internet di Indonesia 63 Juta Orang,” 2016. https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker.

[3] Y. Hadiyat, “Pola komunikasi prostitusi daring di Twitter,” *J. PIKOM (Penelitian Komun. dan Pembangunan)*, vol. 18, no. 2, pp. 125–136, 2017.

[4] W. A. Mukti, “Analisa dan perbandingan bukti forensik aplikasi media sosial facebook dan twitter pada smartphone android.” Fakultas Sains Dan Teknologi Universitas Islam Negeri Syarif Hidayatullah …, 2017.

[5] W. J. S. Poerwadarminta, “Kamus Umum Bahasa Indonesia, Jakarta: Balai Pustaka, 1991, cet.” XII.

[6] B. Simandjuntak, *Patologi Sosial*. Bandung: TARSITO, 1985.

[7] J. Moedjahedy, “Forensik Komputer Studi Kasus: Universitas Klabat,” *E-JURNAL JUSITI J. Sist. Inf. dan Teknol. Inf.*, vol. 5, no. 2, pp. 96–106, 2016.

[8] I. Riadi and R. Umar, “Identification Of Digital Evidence On Android ’ s,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 15, no. 5, pp. 3–8, 2017.

[9] F. Albanna and I. Riadi, “Forensic Analysis of Frozen Hard Drive Using Static Forensics Method,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 15, no. 1, 2017.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Analisis Keamanan Aplikasi Android Dengan Metode Vulnerability Assessment

I Kadek Aldy Oka Ardita[a1], I Gusti Ngurah Anom Cahyadi Putra[a2], Mohammad Rizky Kustiadie [a3], I Gusti Ngurah Made Dika Varuna [a4], Made Yayang Eka Prananda [a5]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Unud, Jimbaran, Bali, 80361, Indonesia
[1]aldy.ardita@gmail.com
[2]anom.cp@unud.ac.id
[3]rizkytegal24@gmail.com
[4]ngurahdika22@gmail.com
[5]yayangp32@gmail.com

Abstract

Seiring berkembangnya beragam aplikasi maka sistem Android haruslah tahan terhadap bebagai serangan malware dengan mengamati izin akses yang diberikan oleh pengguna. Penyerang dapat menggunakan kerentanan dalam aplikasi untuk mencuri berbagai informasi penting. Informasi merupakan aset penting dan berharga berupa rekaman suara, rekaman video, catatan, dll. Oleh karena itu, diperlukan suatu analisis keamanan. dari aplikasi yang digunakan dengan tes / tindakan pada tingkat keamanan aplikasi. Dalam melakukan Vulnerability test atau proses identifikasi celah keamanan pada aplikasi android dilakukan dua teknik yaitu dengan MobSF dan dengan frida. Hasil dari Analisis MobSF sangat terlihat perbedaannya antara mendownload aplikasi melalui pihak ketiga dengan mendownload aplikasi melalui Play Store. Dimana nilai hash yang didapat sangat berbeda baik dari md5, sha1, atau sha256, dari hasil tersebut dapat diketahui bahwa ada perubahan pada file yang disediakan oleh penyedia pihak ketiga. Pada security score didapatkan bahwa aplikasi yang di download melalui pihak ketiga terdapat banyak server dan aktivitas mencurigakan, sedangkan aplikasi yang terdapat di playstore terdapat 2 server yang asli. Pada size, ukuran file yang disediakan oleh pihak ketiga, ukuran file asli hanya 20.37MB sedangkan file yang di sediakan oleh pihak ketiga berukuran 61.33MB. Pada analisis menggunakan frida dilakukan proses penyerangan yaitu bypass login. Dimana pada aplikasi pihak ketiga sudah memiliki email yang telah diinputkan oleh penyedia aplikasi.

Dari hasil analisis yang dilakukan maka lebih baik untuk mendownload aplikasi melalui playstore agar lebih aman. Karena sebagai pengguna awam tidak akan tahu perubahan file apa yang dilakukan dan beresiko atau tidaknya perubahan tersebut terhadap perangkat tersebut

Keywords: *Mobile app, Keamanan, Vulnerability*

1. Pendahuluan

Saat berbagai aplikasi dikembangkan, sistem Android harus tahan terhadap berbagai serangan malware dengan memperhatikan izin akses yang diberikan oleh pengguna. Penyerang dapat menggunakan pelanggaran keamanan dalam aplikasi untuk mencuri informasi[1]. salah satu aset penting dan berharga berupa rekaman suara, rekaman video, catatan, dll. Oleh karena itu, perlu adanya analisis keamanan terhadap aplikasi yang digunakan dengan cara menguji/mengukur tingkat keamanan suatu aplikasi.

Selain menjadi sistem operasi ponsel yang paling populer, Android merupakan sistem operasi yang paling rentan. Banyaknya peretas yang memanfaatkan celah dalam sistem dan aplikasi pihak ketiga [2]. Menurut laporan para peneliti dari TheBestVPN menghitung jumlah kerentanan yang ada pada platform Linux, Windows, dan Android. Hasilnya, sistem keamanan yang dimiliki oleh Google yaitu Android menjadi sistem operasi yang menduduki peringkat pertama dalam jumlah kerentanan pada 2019, dengan total 414 kerentanan. Namun, jumlah kerentanan tersebut mengalami penurunan dari tahun ke tahun. Google memiliki keterbukaan pada sistem operasinya sehingga dapat mengetahui

seberapa rentan sistem operasi yang dimilikinya. Sehingga membuat Google terus memperbaiki sistem operasinya yaitu Android agar tidak mudah di retas .
Beberapa cara telah dilakukan oleh peneliti untuk memperbaiki sebuah sistem keamananan yang ada pada sistem operasi Android. Salah satunya yaitu dilakukannya penetration test [3]. Penetration test adalah mensimulasikan sebuah serangan yang dilakukan karena adanya kerentanan dan menganalisis kerentanan tersebut [4].
Dikarenakan kerentanan yang ada pada sistem operasi Android maka, pada penelitian kali ini akan melakukan analisa keamanan aplikasi Android dengan menggunakan penetration test yang akan menguji ketahanan dari sebuah aplikasi Android. Hasil analisa keamanan aplikasi Android dengan penetration test diharapkan dapat memberi kesadaran bagi pengguna aplikasi dan juga memberikan saran kepada pihak pengembang aplikasi untuk selalu meningkatkan keamanan serta dapat menyadarkan pengguna akan risiko keamanan yang ada pada setiap aplikasi, terdapat dua metode analisis kerentanan aplikasi android diantaranya adalah analisis statis dan analisis dinamis [5], dalam penelitian ini penulis menggunakan metode statis dan dinamis untuk melakukan uji, uji kerentanan sistem secara statis akan menggunakan *tools* bawaan kali linux yaitu MobSf, *tolls* ini merupakan Mobile Security Framework (MobSF) adalah tool otomatis scanning yang biasa digunakan untuk analisis malware, dan security assessment framework yang mampu melakukan analisis statis dan dinamis [6], tools ini digunakan karena memiliki tingkat akses yang sangat mudah. Analisis dinamik akan menggunakan *tools* bawaan kali linux yang bernama frida, tools ini digunakan kerana dimungkinkan menggunakan emulator android yang terkoneksi menggunakan *virtual* usb.

2. Metode

Meote yang akan digunakan pada penelitian ini adalah metode kualitatif yang akan berfokus pada analisis keamanan suatu aplikasi mobile, penelitian ini akan dilaksanakan dengan memanfaatkan *tools* bawaan dari sistem oprasi kali linux.

2.1. Analisis kebutuhan

Kebutuhan non-fungsional:

a. Hardware (perangkat keras)
   1. SSD 512 GB
   2. Ram 8 GB
   3. Intel Core i5
   4. Geforce GTX 1650Ti

b. Software (perangkat lunak)
   1. Kali Linux
   2. Mobile Security Framework (MobSF).
   3. Frida
   4. VMWare

Kebutuhan fungsional:

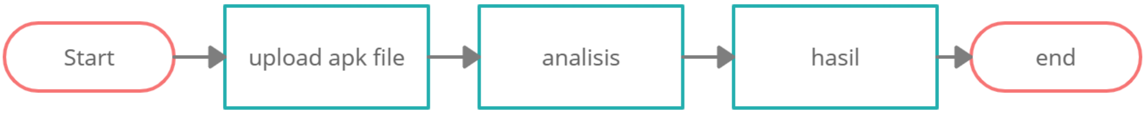
a. Kemampuan ekstrak .apk file
b. Kemampuan analisis celah keamanan

2.2. Rancangan analisis sistem

Dalam penelitian ini, peneliti menerapkan dua metode dalam melakukan *Vulnerability test* atau proses identifikasi celah keamanan pada aplikasi android, yaitu:

a. Analisis statis menggunakan MobSF

Dalam melakukan analisis statik akan digunakan software Mobile Security Framework (MobSF), tahapan yang dilakukan dalam uji statik dapat dilihat pada flowchart pada gambar 1.

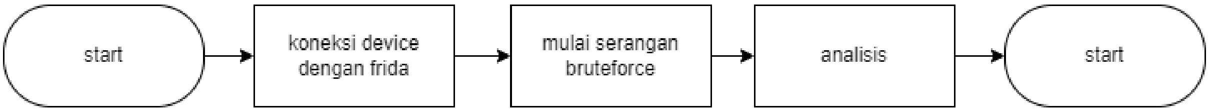


Gambar 1. flowchart pengujian statis

Tahapan awal adalah melakukan persiapan seperti menentukan aplikasi yang akan di uji, pada penelitian ini akan digunakan aplikasi android yang cukup populer yaitu spotify, pada tahapan persiapan penulis mendownload aplikasi spotify dari *play store* dan juga aplikasi spotify yang ada pada link berikut: https://www.goapkmods.com/apps/spotify-premium/. Setelah aplikasi berhasil didapatkan proses akan dilanjutkan dengan proses upload. Setelah proses upload file selesai akan didapatkan beberapa hasil diantaranya adalah informasi mengenai hash sum aplikasi, meta data aplikasi, dan file source code yang dapat digunakan untuk melakukan analisis manual terhadap code.

b. Analisis dinamis menggunakan frida

Frida adalah salah satu tools yang dijalankan menggunakan linux dengan tujuan melakukan analisis dinamik terhadap aplikasi android, berbeda dengan MobSF yang dimana MobSF adalah tools yang digunakan untuk melakukan analisis statik pada sebuah aplikasi android. Langkah - langkah yang dilakukan dalam menggunakan tools ini adalah langkah pertama adalah mengkoneksikan perangkat dengan tools, selanjutnya melakukan beberapa serangan terhadap aplikasi yang telah ditargetkan, terdapat beberapa serangan yang dapat dilakukan salah satunya adalah brute force, flowchar dapat dilihat pada gambar 2.



Gambar 2. flowchart pengujian dinamis

3. Hasil dan pembahasa

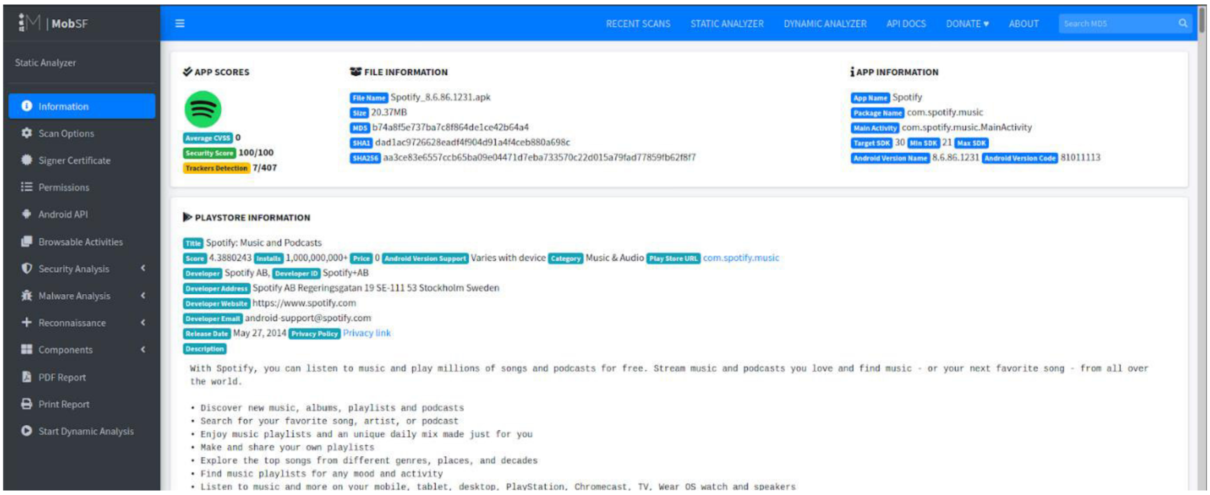
Pada tahapan analisi akan dijelaskan hasil analisis statiks dan dinamis yang telah dilaksanakan.

3.1. Analisis statis menggunakan MobSF

Analisis statis akan dilakukan dengan melakukan upload file apk dan membandingan hasil analisis dari MobSF, terdapat beberapa perbedaan yang mencolok pada hasil kedua aplikasi tersebut perbandingan hasil dapat dilihat pada tabel 1 dan Informasi hasil analisis tersebut dapat dilihat pada gambar 2 dan gambar 3

Table 1. Perbandingan hasil analisis

No	Informasi	Spotify asli	Spotify mod
1	*size*	20.37 MB	Nama: dec12.png
2	*MD5*	b74a8f5e737ba7c8f864de1ce42b64a4	6996e20236ee2649cab6cb4e451vd73e
3	*Security score*	100 / 100	35 / 100
4	*Trackers detection*	7 /407	7 / 407



APP SCORES

Average CVSS 0
Security Score 100/100
Trackers Detection 7/407

FILE INFORMATION

File Name Spotify_8.6.86.1231.apk
Size 20.37MB
MD5 b74a8f5e737ba7c8f864de1ce42b64a4
SHA1 dad1ac9726628eadf4f904d91a4f4ceb880a698c
SHA256 aa3ce83e6557ccb65ba09e04471d7eba733570c22d015a79fad77859fb62f8f7

APP INFORMATION

App Name Spotify
Package Name com.spotify.music
Main Activity com.spotify.music.MainActivity
Target SDK 30 Min SDK 21 Max SDK
Android Version Name 8.6.86.1231 Android Version Code 81011113

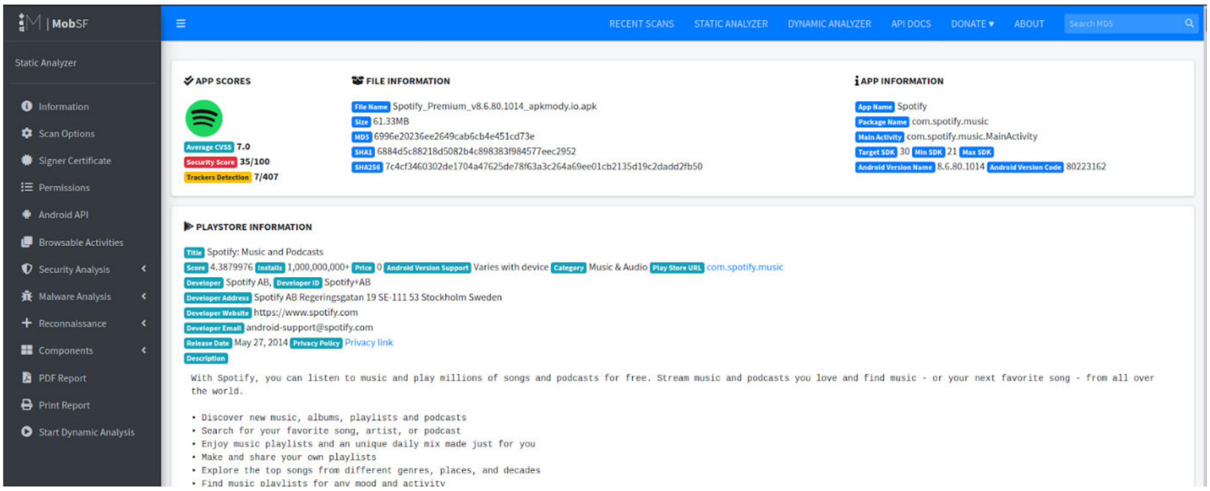
PLAYSTORE INFORMATION

Title Spotify: Music and Podcasts
Score 4.3880243 Installs 1,000,000,000+ Price 0 Android Version Support Varies with device Category Music & Audio Play Store URL com.spotify.music
Developer Spotify AB, Developer ID Spotify+AB
Developer Address Spotify AB Regeringsgatan 19 SE-111 53 Stockholm Sweden
Developer Website https://www.spotify.com
Developer Email android-support@spotify.com
Release Date May 27, 2014 Privacy Policy Privacy link
Description

```
With Spotify, you can listen to music and play millions of songs and podcasts for free. Stream music and podcasts you love and find music - or your next favorite song - from all over
the world.

• Discover new music, albums, playlists and podcasts
• Search for your favorite song, artist, or podcast
• Enjoy music playlists and an unique daily mix made just for you
• Make and share your own playlists
• Explore the top songs from different genres, places, and decades
• Find music playlists for any mood and activity
• Listen to music and more on your mobile, tablet, desktop, PlayStation, Chromecast, TV, Wear OS watch and speakers
```

Gambar 2. Hasil analisis aplikasi original



APP SCORES

Average CVSS 7.0
Security Score 35/100
Trackers Detection 7/407

FILE INFORMATION

File Name Spotify_Premium_v8.6.80.1014_apkmody.io.apk
Size 61.33MB
MD5 6996e20236ee2649cab6cb4e451cd73e
SHA1 6884d5c88218d5082b4c898383f984577eec2952
SHA256 7c4cf3460302de1704a47625de78f63a3c264a69ee01cb2135d19c2dadd2fb50

APP INFORMATION

App Name Spotify
Package Name com.spotify.music
Main Activity com.spotify.music.MainActivity
Target SDK 30 Min SDK 21 Max SDK
Android Version Name 8.6.80.1014 Android Version Code 80223162

PLAYSTORE INFORMATION

Title Spotify: Music and Podcasts
Score 4.3879976 Installs 1,000,000,000+ Price 0 Android Version Support Varies with device Category Music & Audio Play Store URL com.spotify.music
Developer Spotify AB, Developer ID Spotify+AB
Developer Address Spotify AB Regeringsgatan 19 SE-111 53 Stockholm Sweden
Developer Website https://www.spotify.com
Developer Email android-support@spotify.com
Release Date May 27, 2014 Privacy Policy Privacy link
Description

```
With Spotify, you can listen to music and play millions of songs and podcasts for free. Stream music and podcasts you love and find music - or your next favorite song - from all over
the world.

• Discover new music, albums, playlists and podcasts
• Search for your favorite song, artist, or podcast
• Enjoy music playlists and an unique daily mix made just for you
• Make and share your own playlists
• Explore the top songs from different genres, places, and decades
• Find music playlists for any mood and activity
```

Gambar 3. Hasil analisis aplikasi pihak ketiga

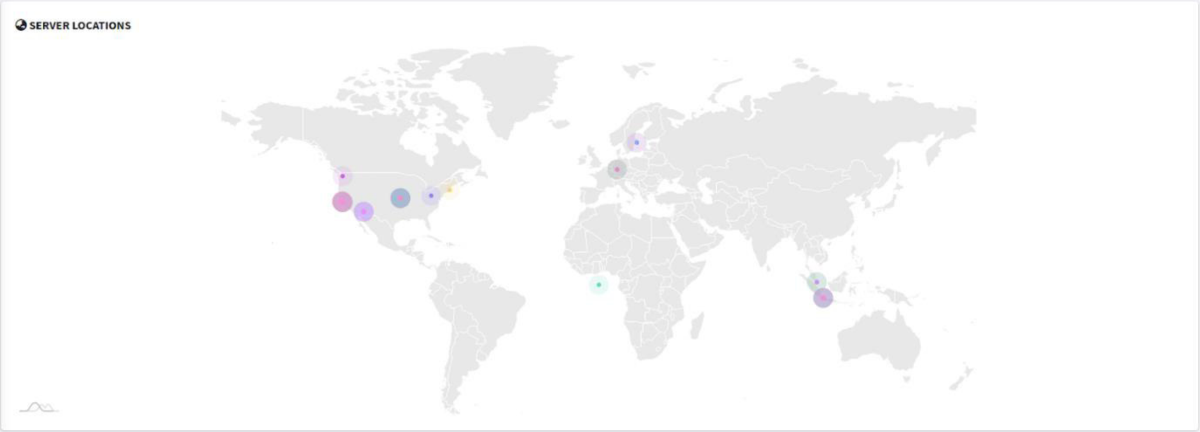
a. Hasil hash

Hashing merupakan salah satu teknik kriptografi yang sering digunakan dalam melakukan identifikasi perubahan pada file atau suatu pesan, dikarenakan hashing memiliki sifat unik yang akan selalu menghasilkan nilai berbeda walau perubahan pada file atau pesan awal sangat minim.

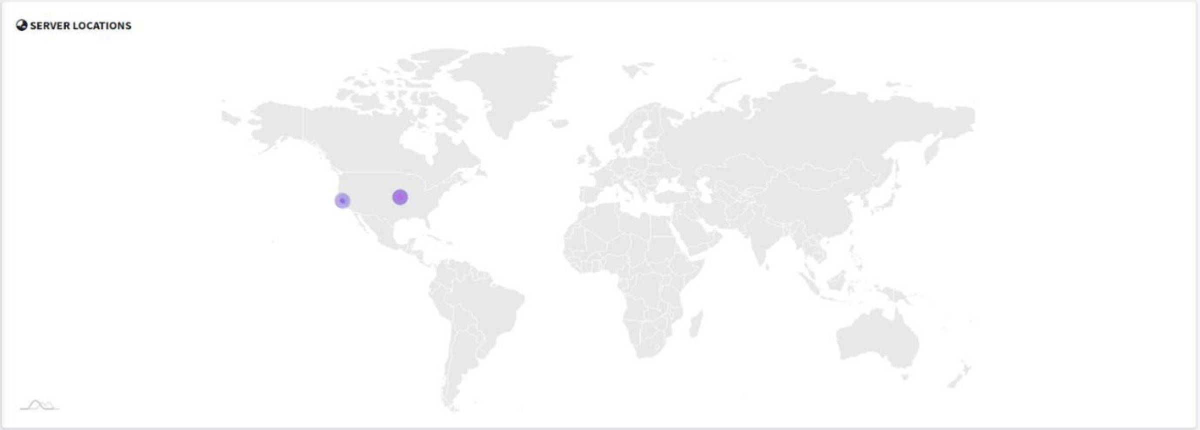
Pada hasil analisis diatas didapatkan bahwa perbedaan hasil hashing baik itu menggunakan md5, sha1, atau sha256 memunculkan nilai yang sangat berbeda dari hasil tersebut dapat diketahui bahwa ada perubahan pada file yang disediakan oleh penyedia pihak ketiga

b. Security Score

Security score merupakan nilai yang diperoleh dengan melakukan analisis terhadap permission dan aktivitas mencurigakan dari aplikasi itu sendiri, pada sistem yang disediakan oleh pihak ketiga memiliki skor keamanan yang rendah dikarenakan aplikasi pihak ketiga memiliki aktivitas yang mencurigakan saat analisis server yang ada, aplikasi original hanya memiliki dua server yang aktif namun aplikasi yang disediakan oleh pihak ketiga memiliki server 10 server yang aktif dan 1 server bekerja secara anonymous, list server dapat dilihat pada gambar 4 dan 5



Gambar 5. List server aplikasi pihak ketiga



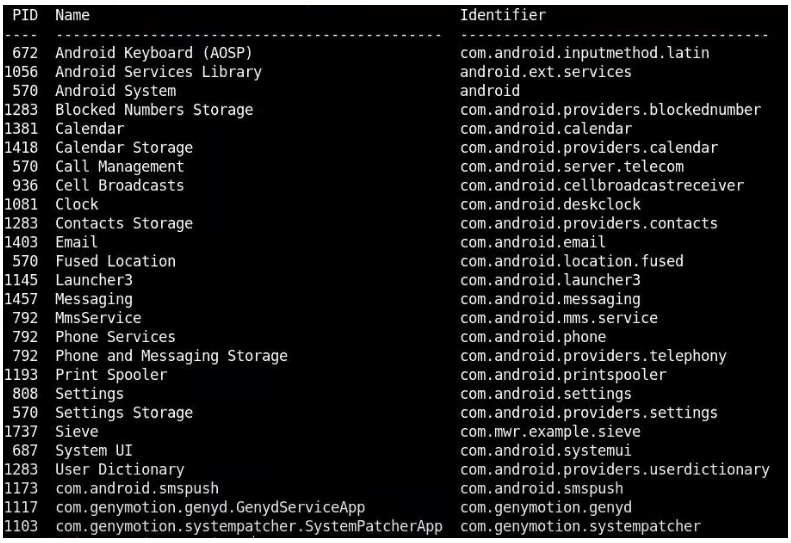
Gambar 6. List server aplikasi original

c. Size

Selain server dan hasil dari hashing juga terdapat perbedaan yang sangat besar pada ukuran file yang disediakan oleh pihak ketiga, ukuran file asli hanya 20.37MB sedangkan file yang di sediakan oleh pihak ketiga berukuran 61.33MB.

3.2. Analisis dinamis menggunakan Frida

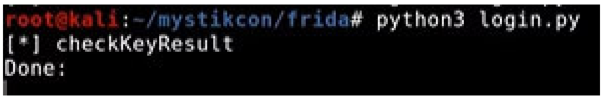
Langkah awal melihat proses yang berjalan pada sistem android



```
 PID  Name                                            Identifier
----  ---------------------------------------------   ------------------------------------
 672  Android Keyboard (AOSP)                         com.android.inputmethod.latin
1056  Android Services Library                        android.ext.services
 570  Android System                                  android
1283  Blocked Numbers Storage                         com.android.providers.blockednumber
1381  Calendar                                        com.android.calendar
1418  Calendar Storage                                com.android.providers.calendar
 570  Call Management                                 com.android.server.telecom
 936  Cell Broadcasts                                 com.android.cellbroadcastreceiver
1081  Clock                                           com.android.deskclock
1283  Contacts Storage                                com.android.providers.contacts
1403  Email                                           com.android.email
 570  Fused Location                                  com.android.location.fused
1145  Launcher3                                       com.android.launcher3
1457  Messaging                                       com.android.messaging
 792  MmsService                                      com.android.mms.service
 792  Phone Services                                  com.android.phone
 792  Phone and Messaging Storage                     com.android.providers.telephony
1193  Print Spooler                                   com.android.printspooler
 808  Settings                                        com.android.settings
 570  Settings Storage                                com.android.providers.settings
1737  Sieve                                           com.mwr.example.sieve
 687  System UI                                       com.android.systemui
1283  User Dictionary                                 com.android.providers.userdictionary
1173  com.android.smspush                             com.android.smspush
1117  com.genymotion.genyd.GenydServiceApp            com.genymotion.genyd
1103  com.genymotion.systempatcher.SystemPatcherApp   com.genymotion.systempatcher
```

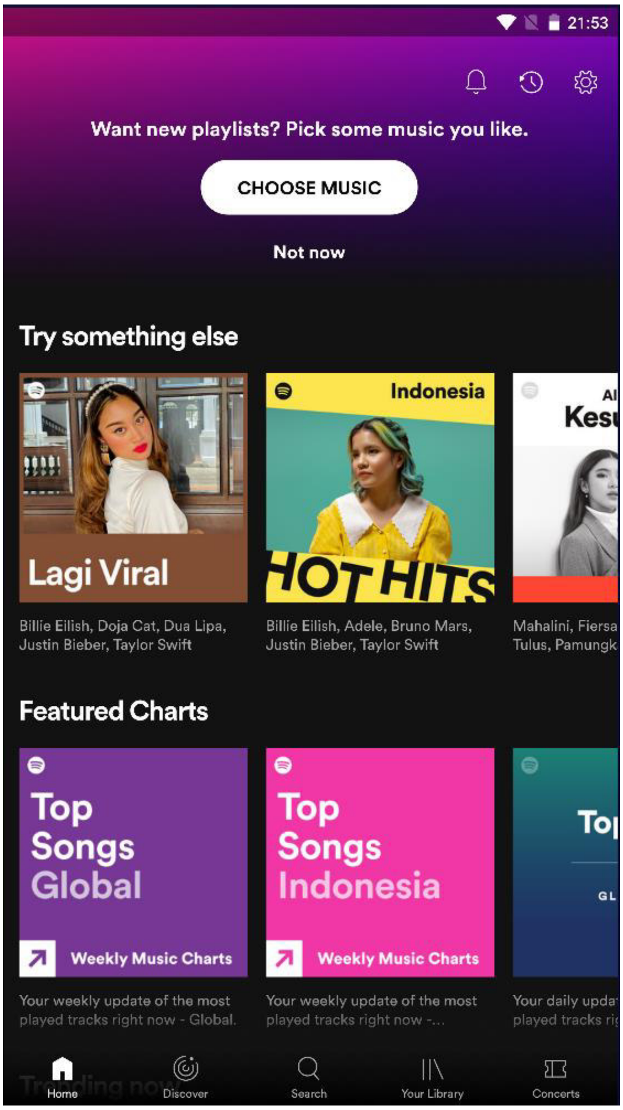
Gambar 7. Aplikasi yang berjalan

Langkah selanjutnya adalah melakukan proses penyerangan, serangan yang di uji cobakan pada aplikasi ini adalah uji coba serangan bypass pada proses login. Sistem yang diujikan memiliki kelemahan terhadap bypass login, pada sistem tidak terjadi begitu banyak perubahan saat aplikasi digunakan tanpa memasukan username dan password hanya sistem masih berjalan normal. Hal ini dikarenakan aplikasi yang disediakan pihak ketiga sudah dimasukan email default oleh penyedia.



```
root@kali:~/mystikcon/frida# python3 login.py
[*] checkKeyResult
Done:
```

Gambar 8. Proses bypass login



21:53
Want new playlists? Pick some music you like.
CHOOSE MUSIC
Not now
Try something else
Lagi Viral
Billie Eilish, Doja Cat, Dua Lipa, Justin Bieber, Taylor Swift
Indonesia
HOT HITS
Billie Eilish, Adele, Bruno Mars, Justin Bieber, Taylor Swift
Mahalini, Fiersa Tulus, Pamungk
Featured Charts
Top Songs Global
Weekly Music Charts
Your weekly update of the most played tracks right now - Global.
Top Songs Indonesia
Weekly Music Charts
Your weekly update of the most played tracks right now -...
Your daily upda played tracks ri
Home
Discover
Search
Your Library
Concerts

Gambar 9. Hasil bypass login

4. Kesimpulan

Dalam melakukan Vulnerability test atau proses identifikasi celah keamanan pada aplikasi android dilakukan dua teknik yaitu dengan MobSF dan dengan frida. Hasil dari Analisis MobSF sangat terlihat perbedaannya antara mendownload aplikasi melalui pihak ketiga dengan mendownload aplikasi melalui Play Store. Dimana nilai hash yang didapat sangat berbeda baik dari md5, sha1, atau sha256, dari hasil tersebut dapat diketahui bahwa ada perubahan pada file yang disediakan oleh penyedia pihak ketiga. Pada security score didapatkan bahwa aplikasi yang di download melalui pihak ketiga terdapat banyak server dan aktivitas mencurigakan, sedangkan aplikasi yang terdapat di playstore terdapat 2 server yang asli. Pada size, ukuran file yang disediakan oleh pihak ketiga, ukuran file asli hanya 20.37MB sedangkan file yang di sediakan oleh pihak ketiga berukuran 61.33MB.

Pada analisis menggunakan frida dilakukan proses penyerangan yaitu bypass login. Dimana pada aplikasi pihak ketiga sudah memiliki email yang telah diinputkan oleh penyedia aplikasi.

Dari hasil analisis yang dilakukan maka lebih baik untuk mendownload aplikasi melalui playstore agar lebih aman. Karena sebagai pengguna awam tidak akan tahu perubahan file apa yang dilakukan dan beresiko atau tidaknya perubahan tersebut terhadap perangkat tersebut.

References

[1] Anwar, Nuril, et al. "Ekstraksi Logis Forensik Mobile pada Aplikasi E-Commerce Android." *Mobile and Forensics* 2.1 (2020): 1-10.

[2] Alviansyah, Fauzan Awanda, and Erika Ramadhani. "Implementasi Dynamic Application Security Testing pada Aplikasi Berbasis Android." *AUTOMATA* 2.1 (2021).

[3] Hanifurohman, Cholis, and Deanna Durbin Hutagalung. "Analisa Keamanan Aplikasi Mobile E-Commerce Berbasis Android Menggunakan Mobile Security Framework." *PROCEEDINGS UNIVERSITAS PAMULANG* 1.1 (2020).

[4] Hanifurohman, Cholis, and Deanna Durbin Hutagalung. "ANALISIS STATIS MENGGUNAKAN MOBILE SECURITY FRAMEWORK UNTUK PENGUJIAN KEAMANAN APLIKASI MOBILE E-COMMERCE BERBASIS ANDROID." *Sebatik* 24.1 (2020): 22-28.

[5] Kartono, Aan, Anang Sularsa, and Setia Juli Irzal Ismail. "Membangun Sistem Pengujian Keamanan Aplikasi Android Menggunakan Mobsf." *eProceedings of Applied Science* 5.1 (2019).

[6] Rama, Gilang Aditya, Fauziah Fauziah, and Nurhayati Nurhayati. "Perancangan Sistem Keamanan Brangkas Menggunakan Pengenalan Wajah Berbasis Android." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 4.3 (2020): 635-641.

[7] Merina, Calysta. *Analisis perbandingan kinerja test automation framework untuk functional testing pada aplikasi berbasis android dengan metode the distance to the ideal alternative*. BS thesis. Fakultas Sains Dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta.

[8] Yumnun, Luqman Hakim, Ari Kusyanti, and Dany Primanita Kartikasari. "Implementasi OWASP Mobile Security Testing Guide (MSTG) Untuk Pengujian Keamanan Pada Aplikasi Berbasis Android." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN* 2548 (2020): 964X.

This page is intentionally left blank.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Comparison of K-Nearest Neighbor And Modified K-Nearest Neighbor With Feature Selection Mutual Information And Gini Index In Informatics Journal Classsification

Benedict Emanuel Sutrisna[a1], AAIN Eka Karyawati[a2], Luh Arida Ayu Rahning Putri[a3], I Wayan Santiyasa[a4], Agus Muliantara [a5], I Made Widiartha[a6],

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
[1]benemanuel0805@gmail.com
[2]eka.karyawati@unud.ac.id
[3]rahningputri@unud.ac.id
[4]santiyasa@unud.ac.id
[5]muliantara@unud.ac.id
[6]madewidiartha@unud.ac.id

Abstract

With the rapid development of informatics where thousands of informatics journals have been made, a new problem has occured where grouping these journals manually has become too difficult and expensive. The writer proposes using text classification for grouping these informatics journals. This research examines the combinations of two machine learning methods, K-Nearest Neighbors (KNN) and Modified K-Nearest Neighbors with two feature selection methods, Gini Index (GI) and Mutual Information (MI) to determine the model that produces the higherst evaluation score. The data are informatics journals stored in pdf files where they are given one of 3 designated labels: Information Retrieval, Database or Others. 252 data were collected from the websites, neliti.com and garuda.ristekbrin.go.id. This research examines and compares which of the two methods, KNN and MKNN at classifying informatics journal as well as determining which combination of parameters and feature selection that produces the best result. This research finds that the combination of method and feature selection that produces the best evaluation score is MKNN with GI as feature selection producing precision score, recall score and f1-score of 97.7%

Keywords: *Text Classification, KNN, MKNN, Mutual Information, Gini Index, Informatics Journal.*

1. Introduction

The field of informatics is experiencing rapid development. Hundreds of research in various fields are conducted each year where their results would be used as material for future research. Though not all findings will be relevant towards a research that's being conducted, as such it would be prudent to group those research to make it easier to find relevant references for future research. Unfortunately, the quick growth of informatics with hundreds of research being published each year makes grouping these research through human efforts near impossible and very expensive. This problem can be overcome with computers through text classifications.

According to [1] various classification methods can be used for document classification in various domains, such as digital libraries and scientific literature. According to [2] one algorithm that can solve the classification problem is K-Nearest Neighbor (KNN) which has an easy to understand and implement algorithm, however it has a weakness where larger dimensionality of data will negatively affect its performance. Several research have been made to overcome this problem, Research conducted by [3] found that feature selection improves the evaluation scores of KNN and Naïve Bayes compared to when both don't use feature selection. [4] created a variant of KNN named Modified K-Nearest Neighbor which has a better evaluation score than KNN. However feature selection was not used in said research, thus it is not known how feature selection would affect its performance. According to [5] the use of the feature selection method Mutual Information (MI) improves the evaluation score of the Support Vector Machine algorithm in classifying Indonesian news articles. [6] found that the Gini Index (GI) feature selection method increases the evaluation score of KNN in classifying cognitive level documents. Based on those sources, the writer believes that both feature selection methods can be used on informatics journals, but wants to know which method produces the highest evaluation score if only use one feature selection method.

Based on the existing problem and the related research of which are the basis of this research, the writer intends to compare KNN and MKNN with MI and GI as feature selection with the hopes that this research find the combination algorithm and feature selection with the highest evaluation score.

2. Reseach Methods

2.1 Research Stage

This research is divided in to two stages. In the first stage, models, which are combinations of algorithms and feature selection methods, are divided in to 3 categories based on which feature selection methods are used, namely: none, GI, and MI. The best model of each category is chosen to continue for the second stage. In the second stage, the 3 chosen models is tested again to determine the best model. Testing in this research is divided in to 2 phases, the training phase and the testing phase. The training phase is where training data is processed so that the model can use it in testing phase. It consists of preprocessing, TF-IDF weighting and feature selection. The testing phase is where the testing data is classified by the model and its results are evaluated. It consists of preprocessing, TF-IDF weighting, feature selection, classification, and evaluation.

2.2 Data Collection

Data is collected from 2 web sources, https://www.neliti.com/id/conferences/semnasif and https://garuda.ristekbrin.go.id/area. 252 information journals were collected and divided evenly in to 3 labels, Information Retrieval, Information and Database Systems and Their Applications, and others. Data labeling is done by the writer and evaluated by 12 fellow students from Text Mining and Big Data disciplines using the Kappa statistic.

2.3 Preprocessing

Preprocessing aims to make the input documents more consistent to facilitate text representation, which is necessary for most text analytics task [1]. As can be seen in Figure 1, this research applies several preprocessing methods, namely case folding, punctuation removal, stemming, stop word removal and tokenization. Case folding is the process of converting letter in to the same case, particularly uppercase letters in to lowercase letters. Stop words removal is the removal of very common and low information words known as stop words. Stemming is the process of cutting inflected words in to their word stem. Tokenization is the process of dividing text in to several units called tokens, the tokens in this research consist of individual words.

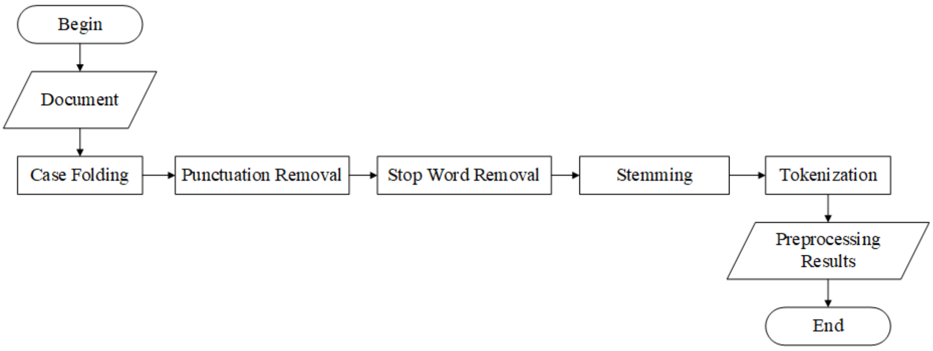


Figure 1. Preprocessing

2.4 TF-IDF Weighting

Term Frequency – Inverse Document Frequency (TF-IDF) is a composite weighting method for each tem in every document. TF-IDF assumes has a good class of distinction occurs if a term has high freqeuncy in one document and low frequency in other documents [2].

p-ISSN: 2301-5373
e-ISSN: 2654-5101

The following are the steps of TF-IDF, see Figure 2:

a. Calculate term frequency of term t in documet d ($tf_{t,d}$)
b. Calculate document frequency of term t (df_t)
c. Calculate inverse document frequency

$$idf_t = log\frac{N}{df_t} \qquad (1)$$

With idf_t as the inverse document frequency of term t, df as the document frequency of term t, and N as the total number of documents

d. Calculate TF-IDF

$$W_{t,d} = tf_{t,d} \times idf_t \qquad (2)$$

With $W_{t,d}$ as weight of term t in document d.

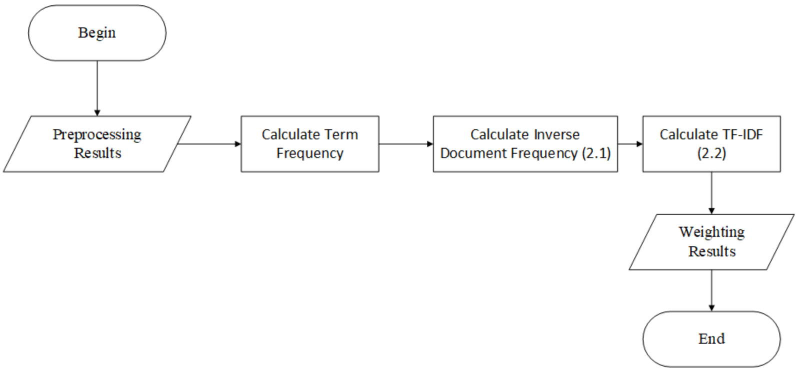


Figure 2. TF-IDF Weighting

2.5 Gini Index

Gini Index (GI) is a measurement of statistical dispersion intended to represent wealth distribution of a country developed by Corrado Gini. GI is often used to measure discriminative power in a feature. GI is typically used for categorical variables, but can be generalized to numeric attributes through discretization [7]. The GI formula is as follows:

$$GI(x) = 1 - \sum_{i=1}^{Y} P(i)^2 \qquad (3)$$

With Y as total labels, x as term, and p(i) as probability of term x in document labeled i.

The steps of Gini Index can be seen in Figure 3.

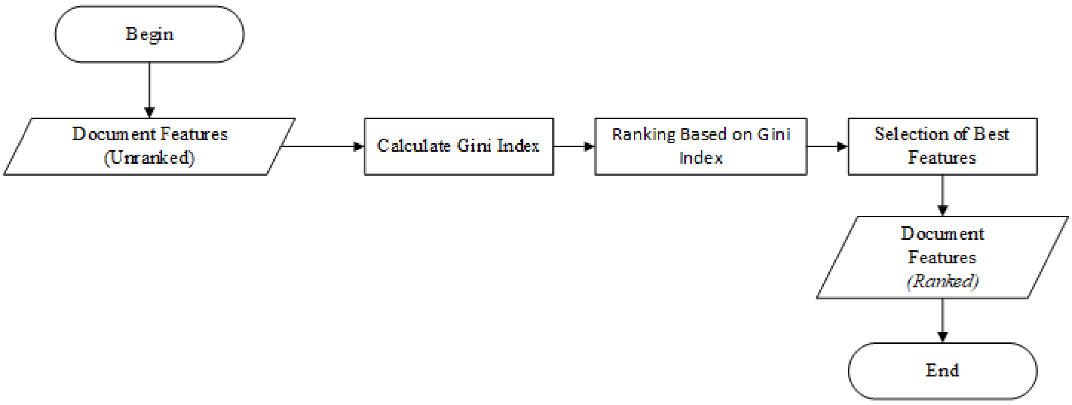

Begin
Document Features (Unranked)
Calculate Gini Index
Ranking Based on Gini Index
Selection of Best Features
Document Features (Ranked)
End

Figure 3. Gini Index

2.6 Mutual Information

According to [8] mutual information (MI) is the measurement of the amount of information that one random variable contains about another random variable. MI is the reduction of uncertainty of a random variable caused by information from another random variable. MI determines the correlation between two words in a data set, if the MI score is large then the two terms often co-occur thus they relate semantically. Conversely a small MI score means that when one of them appears then the other does not, indicating no semantic relation. The formula for MI is as follows:

$$I(x,y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (4)$$

With p(x,y) as joint probability of x and y, p(x) as probability of x, and p(y) as probability of y.

The steps of Mutual Information can be seen in Figure 4.

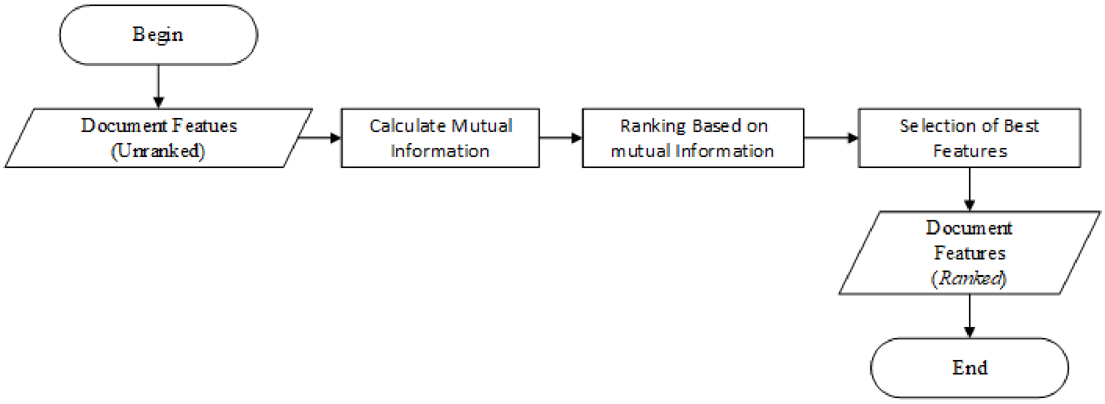

Begin
Document Featues (Unranked)
Calculate Mutual Information
Ranking Based on mutual Information
Selection of Best Features
Document Features (Ranked)
End

Figure 4. Mutual Information

2.7 K-Nearest Neighbours

K-nearest neighbours (KNN) locally determines the decision boundary (label). For 1NN, each document is inserted in to the label of its nearest neighbours. For KNN, each document is inserted in to the majority label of its k nearest neighbours, with k as a parameter. KNN classification is based on contiguity hypothesis, which assumes a document d has the same class as its neighbouring training document [2]. The following are the steps of KNN classification, with the flowchart shownin Figure 5:

a. Determine the value of k.
b. Calculate the distance of the object with each data point. Calculation is done using Euclidian distance with the following formula:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

With D as distance, and x and y as training data and testing data respectively.
c. Gather the data points with the smallest distance as many as k.
d. Determine the class majority of the gathered data points.

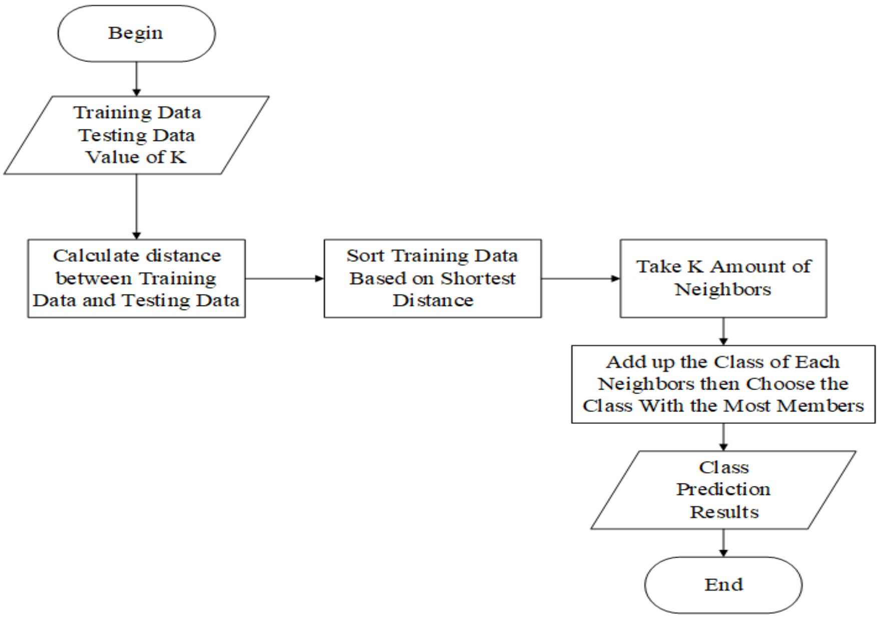

Begin
Training Data
Testing Data
Value of K
Calculate distance between Training Data and Testing Data
Sort Training Data Based on Shortest Distance
Take K Amount of Neighbors
Add up the Class of Each Neighbors then Choose the Class With the Most Members
Class Prediction Results
End

Figure 5. K-Nearest Neighbor

2.8 Modified K-Nearest Neighbours

According to [4] Modified K-Nearest Neighbours (MKNN) is a variation of KNN which computes a kind of weight named validity on training data based on the number of same labeled neighbours divided by the total of neighbors. The following is the algorithm of MKNN, with the flowchart shown in Figure 6:

a. Determine the value of K.

b. Determine validity (v) for each training data with the formula:

$$v(x) = \frac{1}{H}\sum_{i=1}^{H} S\left(lbl(x), lbl\left(N_i(x)\right)\right) \quad (6)$$

With the function S to calculate similarity between x and the i_{th} nearest neighbour with the formula:

$$S(a,b) = \begin{cases} 1 a = b \\ 0 a \neq b \end{cases} \quad (7)$$

With H as the number of neigbors to calculate v, x as designated training data, lbl(x) as label of x, and $N_i(x)$ as i_{th} nearest neighbor of x

c. Calculate the weight of k nearest neighbor with the formula:

$$W(i) = v(i) \times \frac{1}{d+0.5} \quad (8)$$

With W(i) as weight of i_{th} neighbour and d as Euclidean distance

d. Compute the sum weights of every neighbour according to their label.

e. Choose the label with the highest total weight.

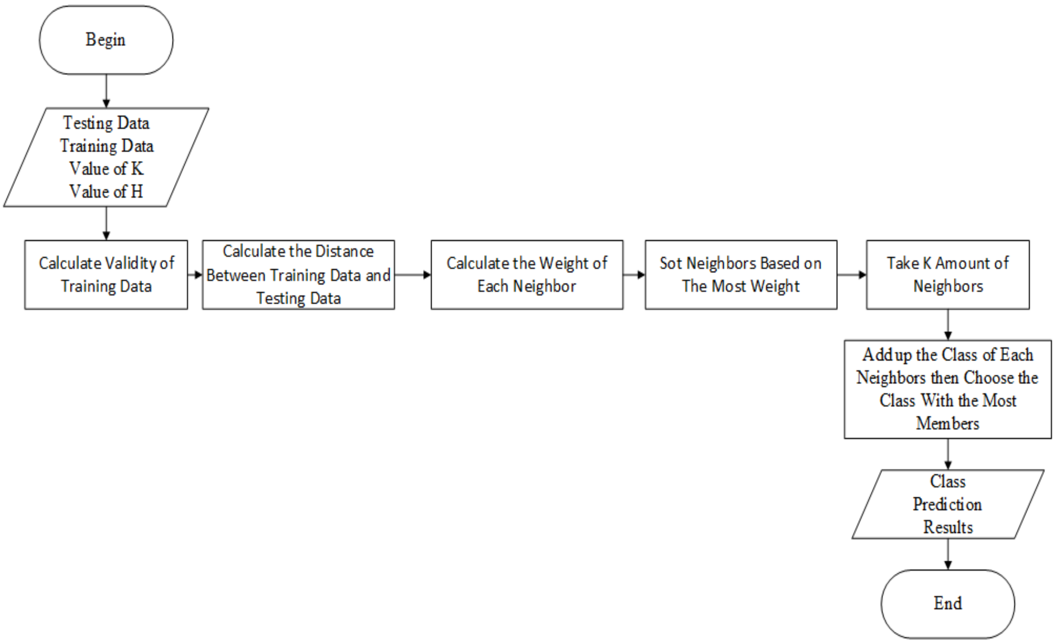

Begin
Testing Data
Training Data
Value of K
Value of H
Calculate Validity of Training Data
Calculate the Distance Between Training Data and Testing Data
Calculate the Weight of Each Neighbor
Sot Neighbors Based on The Most Weight
Take K Amount of Neighbors
Add up the Class of Each Neighbors then Choose the Class With the Most Members
Class Prediction Results
End

Figure 6. Modified K-Nearest Neighbor

2.9 Evaluation

Measurement of each model's effectiveness in classification is done by using precision, recall and f1-score as evaluation scores. Precision is the ratio of total true positive to the sum total of true positive and false positive prediction. Recall is the ratio of total true positive to the sum total true positive and false negative prediction. F1-score is a calculation that combines precision and recall. The formula of precision, recall and f1-score is as the following.

$$Precision = \frac{TTP}{TTP+TFP} \quad (9)$$

$$Recall = \frac{TTP}{TTP+TFN} \quad (10)$$

$$F1 = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \quad (11)$$

With:

- TTP is the total true positive prediction
- TFP is the total false positive
- TFN is the total false negative

The precision, recall and f1-score of each model is recorded and compared with emphasis on f1-score for deciding the best model.

3. Result and Discussion

3.1 Choosing the best model of each category

The following are comparisons between KNN and MKNN with various parameters in 3 categories, without feature selection, with GI, and with MI. The best model of each category will be compared in the next round of testing.

Comparison of Models without Feature Selection

Table 1 is the evaluation result of KNN and MKNN without feature selection. The testing finds that KNN with the parameters k = 3 produced the best result with an f1-score of 25.1%

Table 1. Evaluation Results of Models without Feature Selection

Metode	Precision (average)			Recall (average)			F1-score (average)		
	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
KNN	**31.4%**	24.4%	28.6%	36.2%	**38.6%**	37.6%	23.1%	**25.1%**	23.4%
MKNN (h=10)	11.1%	11.1%	11.1%	33.3%	33.3%	33.3%	16.7%	16.7%	16.7%
MKNN (h=20)	11.1%	11.1%	11.1%	33.3%	33.3%	33.3%	16.7%	16.7%	16.7%
MKNN (h=30)	11.1%	11.1%	11.1%	33.3%	33.3%	33.3%	16.7%	16.7%	16.7%

Comparison of Models with GI

Table 2 is the evaluation result of KNN and MKNN with GI as feature selection. The testing finds that MKNN with the parameters k = 3 and h = 30 produced the best result with an f1-score of 95.5%.

Table 2. Evaluation Results of Models with GI

Metode	Precision (average)			Recall (average)			F1-score (average)		
	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
KNN	95.0%	93.8%	93.8%	94.9%	93.7%	93.6%	94.9%	93.3%	93.3%
MKNN (h=10)	91.7%	90.5%	90.3%	90.4%	89.5%	89.4%	90.2%	89.0%	88.9%
MKNN (h=20)	93.6%	92.4%	92.2%	94.3%	93.3%	93.3%	94.0%	92.4%	92.2%
MKNN (h=30)	**95.8%**	95.2%	95.2%	**95.4%**	94.8%	94.7%	**95.5%**	94.8%	94.7%

Comparison of Models with MI

Table 3 is the evaluation result of KNN and MKNN with MI as feature selection. The testing finds that MKNN with the parameters k = 3 and h = 20 produced the best result with an f1-score of 91.3%.

Table 3. Evaluation Results of Models with MI

Metode	Precision (average)			Recall (average)			F1-score (average)		
	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
KNN	91.9%	88.6%	87.8%	90.6%	85.7%	84.2%	89.3%	82.9%	80.6%
MKNN (h=10)	92.9%	89.5%	88.9%	**92.1%**	88.6%	88.0%	90.5%	85.2%	84.0%
MKNN (h=20)	**93.0%**	90.0%	89.4%	91.9%	88.1%	87.1%	**91.3%**	86.7%	85.3%
MKNN (h=30)	91.7%	88.1%	87.3%	91.2%	87.1%	86.1%	90.8%	85.7%	84.3%

3.2 **Comparison Between Models**

This section compares the best models chosen in section 3.1. Table 4 is the evaluation result from testing KNN with k = 5 and no feature selection (model 1). Table 5 is the evaluation result from testing MKNN with k = 3, h = 30 and GI as feature selection (model 2). Table 6 is the evaluation result from testing MKNN with k = 3, h = 20 and MI as feature selection (model 3). From the testing of the three models, model 2 produced the best result with an average f1-score of 97.7%, followed by model 3 producing an average f1-score of 95%, and finally model 1 produced an average f1-score of 30% which is the worst of the results.

Table 4. Testing Results of Model 1

	Precision	Recall	F1-score
IR	0.0%	0.0%	**0.0%**
DB	38.0%	100.0%	**55.0%**
Other	75.0%	23.0%	**35.0%**
Average	37.7%	41.0%	**30.0%**

Table 5. Testing Results of Model 2

	Precision	Recall	F1-score
IR	100.0%	93.0%	**96.0%**
DB	93.0%	100.0%	**97.0%**
Other	100.0%	100.0%	**100.0%**
Rata-rata	97.7%	97.7%	**97.7%**

Table 6. Testing Results of Model 3

	Precision	Recall	F1-score
IR	100.0%	100.0%	**100.0%**
DB	88.0%	100.0%	**93.0%**
Other	100.0%	85.0%	**92.0%**
Rata-rata	96.0%	95.0%	**95.0%**

4. Conclusion

This research found that in classifying informatics journals the best combination of algorithm and feature selection method is MKNN with parameters k = 3 and h = 30 with GI as feature selection, producing an average f1-score of 97.7%. It is worth noting that MKNN with MI as feature selection also produced good results with an average f1-score of 95%. Meanwhile both KNN and MKNN without feature selection scored poorly, the highest score that could be produced being an average f1-score of 30%. In conclusion, the best method to classify informatics journals is MKNN with a feature selection method, preferably GI, but MI is also capable of producing satisfying results.

References

[1] C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*. Boston, MA: Springer US, 2012.

[2] C. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009.

[3] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, “Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm,” *International Journal of Computer Science and Network Security*, vol. 18, no. 12, pp. 95–101, Dec. 2018.

[4] H. Parvin, H. Alizadeh, and B. Minaei-Bidgoli, “MKNN: Modified K-Nearest Neighbor,” in *Proceedings of the World Congress on Engineering and Computer Science 2008*, San Francisco, USA, Oct. 2008, pp. 831–834.

[5] L. G. Irham, A. Adiwijaya, and U. N. Wisesty, “Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine,” *mib*, vol. 3, no. 4, pp. 284–292, Oct. 2019.

[6] T. Setiyorini and R. T. Asmono, “Penerapan Gini Index dan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kognitif Soal Pada Taksonomi Bloom,” *Jurnal Pilar Nusa Mandiri*, vol. Vol. 13, no. 2, pp. 209–216, Sep. 2017.

[7] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015.
[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2006.

This page is intentionally left blank.

Rancang Bangun Portal Lowongan Pekerjaan Berbasis Web Dengan Fitur Rekomendasi

Ida Bagus Gede Dwidasmara[a1], I Putu Yuda Juniantara Putra[a2], I Putu Gede Hendra Suputra[b3], Agus Muliantara[b4], I Gusti Agung Gede Arya Kadyanan[b5], I Wayan Supriana[b6]

[a]Program Studi Teknik Informatika, Jurusan Ilmu Komputer, Fakultas Matermatika dan Ilmu Pengetahuan Alam Universitas Udayana
Jalan Raya Kampus Unud, Badung, 08361, Bali, Indonesia
[1] dwidasmara@unud.ac.id
[2]yudajuniantara@gmail.com
[3]hendra.suputra@unud.ac.id
[4]muliantara@unud.ac.id
[5]dewabayu@unud.ac.id
[6] wayan.supriana@unud.ac.id

ABSTRAK

Tingginya tingkat pengangguran di Indonesia yang disebabkan oleh dampak pertumbuhan penduduk yang padat dan sulitnya mencari pekerjaan yang mengakibatkan pertumbuhan ekonomi tidak stabil menjadi tantangan bagi pemerintah untuk menanggulangi masalah pengangguran. Salah satu faktor yang mempengaruhi tingkat pengangguran yang tinggi adalah penyebaran informasi lowongan kerja yang kurang merata. Pelamar kerja sering kesulitan untuk mendapatkan informasi pekerjaan yang sesuai dengan kemampuan dan keterampilan yang dimiliki. Dengan dibuatnya sistem rekomendasi lowongan kerja, diharapkan dapat menyelesaikan masalah yang dialami seorang pelamar yang ingin melamar pekerjaan yang sesuai dengan kemampuannya. Dengan menggunakan algoritma euclidean distance, sistem diharapkan dapat memberikan hasil rekomendasi berdasarkan beberapa parameter yang sesuai dengan keinginan pelamar. Parameter yang dimaksudkan adalah usia, jenis kelamin, pendidikan, jurusan, gaji, keterampilan, pengalaman kerja, dan kategori pekerjaan yang diinginkan. Penelitian ini menghasilkan kesimpulan yaitu berdasarkan hasil pengujian *precision recall* sistem rekomendasi lowongan kerja menggunakan metode euclidean distance menghasilkan nilai *precision* sebesar 0,94 atau 94% dan nilai *recall* sebesar 0,94 atau 94%.

Kata kunci: Euclidean Distance, Lowongan Kerja, Sistem Rekomendasi Lowongan Kerja

1. Pendahuluan

Tingginya tingkat pengangguran di Indonesia yang disebabkan oleh dampak pertumbuhan penduduk yang padat dan sulitnya mencari pekerjaan yang mengakibatkan pertumbuhan ekonomi tidak stabil menjadi tantangan bagi pemerintah untuk menanggulangi masalah pengangguran. Salah satu faktor yang mempengaruhi tingkat pengangguran yang tinggi adalah penyebaran informasi lowongan kerja yang kurang merata. Masyarakat sering kesulitan untuk mendapatkan pekerjaan yang sesuai dengan kemampuan dirinya karena proses pencarian lowongan kerja, pengajuan Curriculum Vitae (CV), dan proses seleksi masih dilakukan secara manual sehingga membutuhkan waktu yang relatif lama.

Seiring dengan perkembangan teknologi informasi, banyak website yang menyajikan lowongan pekerjaan. Beberapa situs yang menyajikan informasi lowongan pekerjaan yaitu situs jobstreet.com, jobsdb.com dan situs - situs pencari kerja lainnya. Akan tetapi mayoritas website tersebut umumnya hanya menyediakan fasilitas input lowongan kerja bagi perusahaan dan fasilitas melamar kerja online bagi para pencari kerja atau pelamar. Aktivitas yang dapat dilakukan oleh pencari dan penyedia pekerjaan dalam menggunakan website cenderung terbatas hanya mendapatkan informasi lowongan pekerjaan. Para pencari kerja perlu mendapakan rekomendasi pekerjaan berdasarkan kedekatan profil atau data diri pencari kerja dengan persyaratan lowongan kerja dari perusahaan untuk dapat membantu memudahkan dalam menentukan pilihan. Penelitian yang dilkaukan oleh Henny Leidiyana dengan judul

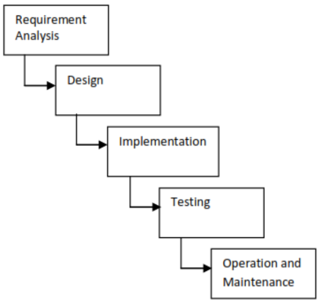
Penerapan Metode K-Nearest Neighbor Pada Penentuan Grade Dealer Sepeda Motor [1]. Pada penelitin tersebut metode pengukuran jarak yang digunakan adalah Euclidean Distance.Berdasarkan permasalahan tersebut, Penelitian ini mengambil judul "Rancang Bangun Portal Lowongan Pekerjaan Berbasis Web Dengan Fitur Rekomendasi". Dengan menerapkan metode Euclidean Distance pada fitur rekomendasi, portal lowongan pekerjaan dapat memberikan rekomendasi berdasarkan kedekatan dari data (profil) pelamar kerja dengan persyaratan dari lowongan pekerjaan.

2. Metode Penelitian

2.1 Pengumpulan Data

Acuan untuk menentukan ukuran data, yaitu minimal 30 sampai dengan 500 data [2]. Penelitian ini menggunakan sumber data yaitu data sekunder yang diperoleh dari website bursakerja.denpasarkota.go.id Dinas Tenaga Kerja dan Sertifikasi Kompetensi Pemerintah Kota Denpasar. Data yang diambil adalah fitur-fitur atau persyaratan lowongan kerja dan contoh data pelamar pekerjaan. Setelah data dikumpulkan kemudian akan dimasukkan ke dalam database.

2.2 Metode Pengembangan Sistem SLDC Waterfall



Gambar 2.1 Tahapan dalam model waterfall (Pfleeger & Atlee, 2010)

Berikut merupakan tahapan-tahapan dalam model waterfall [3]:

1. Requirement Analysis
   Mengumpulkan informasi kebutuhan atas sistem baru, menganalisa dan menyiapkan dokumentasi sistem yang tepat untuk membantu proses pengembangan lebih lanjut. Tahap ini menghasilkan dokumen yang berisi kebutuhan sistem baik kebutuhan fungsional maupun non-fungsional yang sudah diidentifikasi.
2. Design
   Informasi yang diperoleh dari tahap sebelumnya dievaluasi untuk selanjutnya merumuskan implementasi yang tepat. Tahap ini merupakan proses perencanaan dan pemecahan masalah sebagai solusi dari sistem berjalan. Perancangan sistem berguna sebagai gambaran secara logika, struktur dan alir data dari kebutuhan sistem. Perancangan sistem menggunakan bantuan beberapa alat diagram diantaranya DFD (Data Flow Diagram), Flowchart (Diagram alir proses), ERD (Entity Relationship Diagram), dan rancangan antarmuka.
3. Implementation
   Pada tahap implementasi, seluruh rancangan dan desain sistem yang sudah dibuat pada tahap sebelumnya diimplementasikan ke dalam bentuk kode program.
4. Testing
   Tahap ini berkaitan dengan pengujian apakah sistem telah memenuhi setiap kebutuhan yang telah dikumpulkan pada tahap sebelumnya. Di tahap ini juga dilakukan pencarian kesalahan atas implementasi sistem serta perbaikan atas kesalahan tersebut.
5. Operation and Maintenance
   Tahap penerapan program meliputi penerapan sistem secara langsung oleh pengguna. Tahap pemeliharaan meliputi kemungkinan sistem memerlukan beberapa modifikasi dan perbaikan. Dengan kata lain pada tahap ini dilakukan persiapan-persiapan atas segala kemungkinan yang akan terjadi pada sistem.

2.5 Algoritma Euclidean Distance

Metode *Euclidean distance* merupakan suatu metode yang digunakan untuk menghitung jarak antara dua objek [4]. Pada penelitian ini metode *Euclidean distance* ini digunakan untuk pengukuran jarak parameter lowongan pekerjaan dengan dengan data diri pencari kerja. Rumus *Euclidean* untuk pengukuran jarak dua objek adalah:

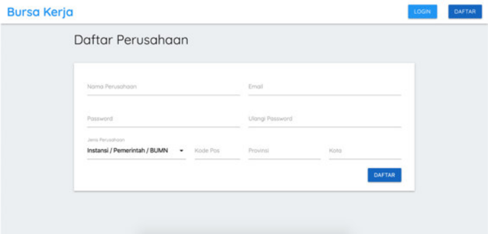
$$d = \sqrt{x^2 + y^2}$$

Jika variable lebih dari dua maka rumus Euclidean untuk pengukuran jarak dua objek adalah:

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

3. Hasil dan Pembahasan

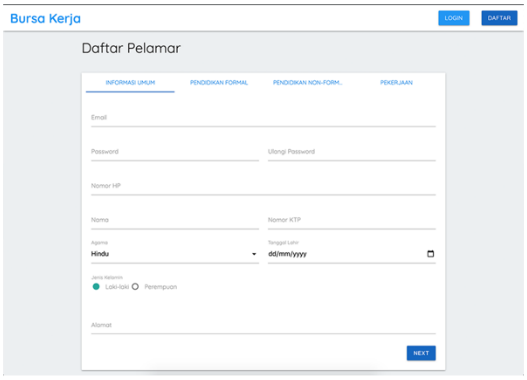
3.1 Implementasi antarmuka sistem



Jenis Perusahaan
Instansi / Pemerintah / BUMN

Gambar 3.1 Tampilan antarmuka daftar perusahaan

Gambar 3.1 merupakan tampilan halaman daftar sebagai perusahaan. Halaman ini muncul setelah pengguna menekan button daftar kemudian pilih daftar sebagai perusahaan. Pada halaman ini terdapat form yang berisi informasi yang harus diisi oleh pengguna untuk mendaftarkan perusahaan. Informasi yang terdapat pada form seperti nama perusahaan, email, password, ulangi password, jenis perusahaan, provinsi, dan kota. Email dan password akan digunakan pengguna untuk login sebagai perusahaan sehingga pengguna dapat mengakses fitur-fitur sebagai perusahaan.



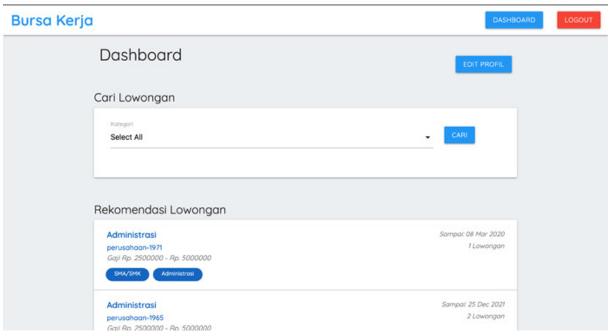
Agama
Hindu

Tanggal Lahir
dd/mm/yyyy

Jenis Kelamin
● Laki-laki ○ Perempuan

Gambar 3.2 Tampilan antarmuka daftar sebagai pelamar

Gambar 3.2 merupakan tampilan antarmuka halaman daftar sebagai pelamar. Halaman ini muncul setelah pengguna menekan button daftar kemudian pilih daftar sebagai pelamar. Pada halaman ini terdapat form yang berisi informasi yang harus diisi oleh pengguna untuk mendaftarkan diri sebagai pelamar. Informasi yang terdapat pada form terbagi kedalam empat bagian yaitu informasi umum, pendidikan formal, pendidikan non-formal, dan pekerjaan.



Dashboard

Cari Lowongan

Kategori: Select All

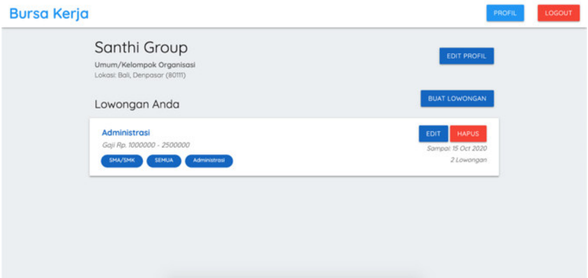
Rekomendasi Lowongan

Administrasi
perusahaan-1971
Gaji Rp. 2500000 - Rp. 5000000
Sampai: 08 Mar 2020
1 Lowongan

Administrasi
perusahaan-1965
Gaji Rp. 2500000 - Rp. 5000000
Sampai: 25 Dec 2021
2 Lowongan

Gambar 3.3 Tampilan antamuka dashboard pelamar

Gambar 3.3 adalah tampilan antarmuka halaman dashboard pelamar. pada fitur ini terdapat dua form yaitu form cari lowongan dan form rekomendasi lowongan. Pada form cari lowongan, pelamar dapat memfilter lowongan berdasarkan pilihan kategori lowongan. Pada bagian form rekomendasi lowongan, pelamar dapat melihat rekomendasi lowongan yang sesuai dengan data yang diisi saat melakukan pendaftaran.



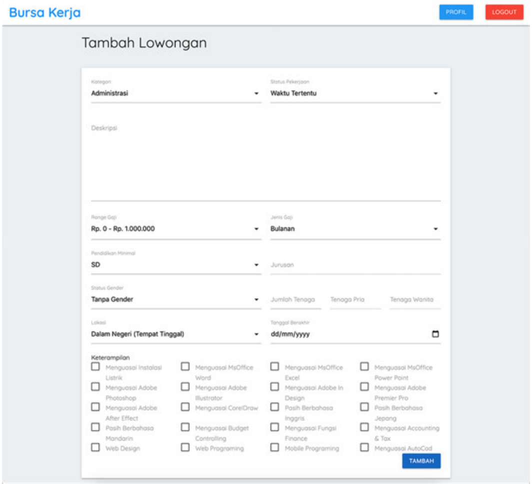
Santhi Group
Umum/Kelompok Organisasi
Lokasi: Bali, Denpasar (80111)

Lowongan Anda

Administrasi
Gaji Rp. 1000000 - 2500000
Sampai: 15 Oct 2020
2 Lowongan

Gambar 3.4 Tampilan antarmuka profil perusahaan

Gambar 3.4 merupakan tampilan antarmuka profil perusahaan. Halaman ini muncul setelah pengguna melakukan login sebagai perusahaan. Pada halaman ini terdapat beberapa informasi seperti nama perusahaan, jenis perusahaan, dan lokasi perusahaan yang dapat dilihat pada bagian kiri sedangkan pada bagian kanan terdapat dua button yaitu edit profil dan buat lowongan.



Tambah Lowongan

Kategori: Administrasi
Status Pekerjaan: Waktu Tertentu
Deskripsi
Range Gaji: Rp. 0 - Rp. 1.000.000
Jenis Gaji: Bulanan
Pendidikan Minimal: SD
Jurusan
Status Gender: Tanpa Gender
Jumlah Tenaga
Tenaga Pria
Tenaga Wanita
Lokasi: Dalam Negeri (Tempat Tinggal)
Tanggal Berakhir: dd/mm/yyyy

Keterampilan
- □ Menguasai Instalasi Listrik
- □ Menguasai Adobe Photoshop
- □ Menguasai Adobe After Effect
- □ Pasih Berbahasa Mandarin
- □ Web Design
- □ Menguasai MsOffice Word
- □ Menguasai Adobe Illustrator
- □ Menguasai CorelDraw
- □ Menguasai Budget Controlling
- □ Web Programing
- □ Menguasai MsOffice Excel
- □ Menguasai Adobe In Design
- □ Pasih Berbahasa Inggris
- □ Menguasai Fungsi Finance
- □ Mobile Programing
- □ Menguasai MsOffice Power Point
- □ Menguasai Adobe Premier Pro
- □ Pasih Berbahasa Jepang
- □ Menguasai Accounting & Tax
- □ Menguasai AutoCad

Gambar 3.5 Tampilan antarmuka tambah lowongan

Gambar 3.5 merupakan tampilan antarmuka halaman tambah lowongan. Halaman ini dapat diakses dengan mengklik button buat lowongan pada halaman profil perusahaan. Pada halaman ini terdapat beberapa form yang harus diisi seperti kategori lowongan, status pekerjaan, range gaji, jenis gaji, pendidikan minimal, jurusan, status gender atau jenis kelamin, jumlah tenaga, lokasi, tanggal berakhir, dan keterampilan. Jika sudah selesai mengisi form dapat dilanjutkan dengan mengklik tambah untuk menyimpan data lowongan baru. Terdapat button *logout* pada bagian kanan atas jika pengguna ingin *logout* dari akun perusahaan.

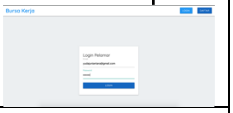
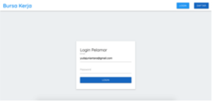
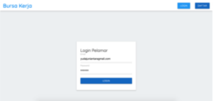
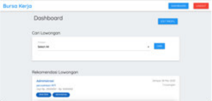
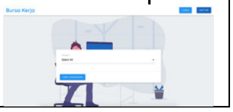
p-ISSN: 2301-5373
e-ISSN: 2654-5101

3.2 Pengujian Sistem

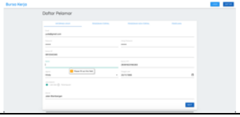
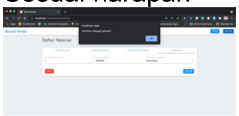
Tahap pengujian merupakan tahap untuk memastikan apakah sistem yang dibuat telah memenuhi tujuan yang ingin dicapai. Pada penelitian ini dilakukan pengujian sistem dengan *black-box testing.* Pengujian sistem dimaksudkan untuk mengetahui sejauh mana tingkat keberhasilan semua tombol dalam tiap menu yang ada didalam sistem seperti menu login, menu pendaftaran, menu rekomendasi dan lain sebagainya. Berikut adalah detail pengujian sistem dengan menggunakan metode *black-box testing*.

3.3 Pengujian Black Box

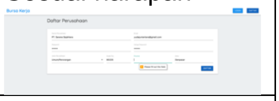
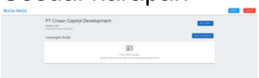
Pengujian Black Box dilakukan untuk mengetahui fungsi spesifik dari aplikasi. Pengujian ini mendemonstrasikan setiap fungsi dari aplikasi dan mengetahui apakah terjadi error atau tidak. Pengujian black box digunakan untuk mengetahui apakah input atau output yang dihasilkan aplikasi sudah sesuai dengan yang diinginkan.

Kode Kebutuhan: **KF1**			
Kasus: ***Pengujian proses login dan logout***			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Pengguna mengisi *input* email dan *password* yang salah	Menampilkan pesan bahwa email atau *password* salah	Sesuai Harapan 
2	Pengguna mengosongkan salah satu *input*	Menampilkan pesan bahwa semua *input* harus diisi	Sesuai harapan 
3	Pengguna mengisi email dengan format yang salah	Menampilkan pesan bahwa format penulisan email salah	Sesuai harapan 
4	Pengguna mengisi email dan *password* yang benar	Pengguna berhasil login dan berpindah ke halaman utama	Sesuai harapan 
5	Pengguna menekan tombol *logout* pada sistem	Pengguna *logout* dari sistem	Sesuai harapan 

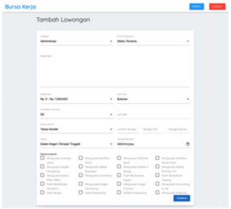
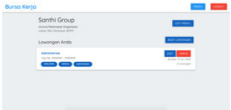
Tabel 3.1Pengujian proses login dan logout

Kode Kebutuhan: **KF2**			
Kasus: **Pengujian proses pendaftaran pelamar kerja**			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Pengguna mengosongkan salah satu *input*	Menampilkan pesan bahwa semua *input* harus diisi	Sesuai harapan 
2	Pengguna mengisi data lengkap dan menekan tombol daftar	Menampilkan pesan pendaftaran berhasil	Sesuai harapan 

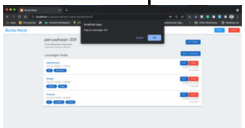
Tabel 3.2 Pengujian proses pendaftaran pelamar kerja

Kode Kebutuhan: **KF3**			
Kasus: **Pengujian proses pendaftaran penyedia kerja**			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Pengguna mengosongkan salah satu *input*	Menampilkan pesan bahwa semua *input* harus diisi	Sesuai harapan  Bursa Kerja Daftar Perusahaan
2	Pengguna mengisi data lengkap dan menekan tombol daftar	Menampilkan pesan pendaftaran berhasil	Sesuai harapan  Bursa Kerja PT Crown Capital Development Lowongan Anda

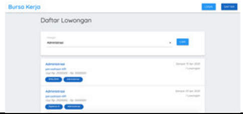
Tabel 3.3Pengujian proses pendaftaran penyedia kerja

Kode Kebutuhan: **KF4**			
Kasus: **Pengujian proses menambah lowongan kerja**			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Pengguna mengosongkan salah satu *input*	Menampilkan pesan bahwa semua *input* harus diisi	Sesuai harapan  Bursa Kerja Tambah Lowongan
2	Pengguna mengisi data lengkap dan menekan tombol tambah lowongan kerja	Menampilkan pesan penambahan data berhasil	Sesuai harapan  Bursa Kerja Santhi Group Lowongan Anda

Tabel 3.4 Pengujian proses menambah lowongan kerja

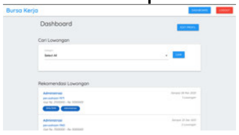
Kode Kebutuhan: **KF5**			
Kasus: **Pengujian proses menghapus lowongan kerja**			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Menghapus data	Menampilkan pesan untuk melakukan pengecekan apakah data yang ingin dihapus sudah benar	Sesuai harapan 

Tabel 3.5 Pengujian proses menghapus lowongan kerja

Kode Kebutuhan: KF6			
Kasus: Pengujian proses mencari lowongan kerja			
No	Skenario Pengujian	Hasil yang Diharapkan	Hasil Pengujian
1	Memilih lowongan kerja berdasarkan kategori	Menampilkan listi lowongan kerja berdasarkan kategori yang dipilih	Sesuai harapan  Bursa Kerja Daftar Lowongan

Tabel 3.6 Pengujian proses mencari lowongan kerja

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Kode Kebutuhan: KF5			
Kasus: Pengujian proses menampilkan rekomendasi lowongan			
No	**Skenario Pengujian**	**Hasil yang Diharapkan**	**Hasil Pengujian**
1	Menampilkan rekomendasi lowongan kerja	Menampilkan list lowongan pekerjaan yang direkomendasikan kepada pelamar	Sesuai harapan  Bursa Kerja Dashboard Cari Lowongan Select All Rekomendasi Lowongan Administrasi Administrasi

Tabel 3.7 Pengujian proses menampilkan rekomendasi lowongan

3.4 Precision and Recall

Precision bersama recall merupakan salah satu pengujian dasar dan paling sering digunakan dalam penentuan efektifitas information retrival system maupun recommendation system. True positive (*tp*) pada information retrival merupakan item relevan yang dihasilkan oleh sistem. Sedangkan false positive (*fp*) merupakan semua item yang dihasilkan oleh sistem. Sehingga dalam information retrival, precision dihitung dengan persamaan berikut [5].

$$Precision = \frac{tp}{tp + fp} = \frac{relevant\ item\ retrieved}{item\ retrieved}$$

$$Recall = \frac{tp}{tp + fn} = \frac{relevant\ item\ retrieved}{relevant\ item}$$

Istilah positive dan negative mengacu pada prediksi yang dilakukan oleh sistem. Sedangkan istilah true dan false mengacu pada prediksi yang dilakukan oleh pihak luar atau pihak yang melakukan observasi. Pembagian kondisi tersebut dapat dilihat pada Tabel [5].

	Relevant	*Non-Relevant*
Retrieved	*True Positif (tp)*	*False Positif (fp)*
Not Retrieved	*False Negative (fn)*	*True Negative (tn)*

Tabel 3.8 Pembagian kondisi hasil yang memungkinkan

No	**Pelamar**	**Jenis Pekerjaan Yang Diinginkan**	**Jumlah Rekomendasi**	**Jumlah Rekomendasi Relevan**	**Precision**	**Recall**
1	Pelamar - 1	Administrasi	9	9	1	1
2	Pelamar - 2	Administrasi	18	18	1	1
3	Pelamar - 3	Perhotelan	2	2	1	1
4	Pelamar - 4	Perhotelan	2	2	1	1
5	Pelamar - 5	Administrasi	18	18	1	1
6	Pelamar - 6	Customer Services	6	6	1	1
7	Pelamar - 7	Perdagangan Besar, Eceran dan Rumah Makan	6	6	1	1
8	Pelamar - 8	Industri Pengolahan	1	1	1	1
9	Pelamar - 9	Perhotelan	2	2	1	1

10	Pelamar - 10	Pendidikan	3	3	1	1
11	Pelamar - 11	Administrasi	6	6	1	1
12	Pelamar - 12	Logistik	2	2	1	1
13	Pelamar - 13	Design	0	0	0	0
14	Pelamar - 14	Rumah Sakit	0	0	0	0
15	Pelamar - 15	Administrasi	6	6	1	1
16	Pelamar - 16	TI (Software)	0	0	0	0
17	Pelamar - 17	Akunting	5	5	1	1
18	Pelamar - 18	Akunting	2	2	1	1
19	Pelamar - 19	Akunting	5	5	1	1
20	Pelamar - 20	Akunting	2	2	1	1
21	Pelamar - 21	Customer Services	4	4	1	1
22	Pelamar - 22	Customer Services	1	1	1	1
23	Pelamar - 23	Customer Services	1	1	1	1
24	Pelamar - 24	Finance	1	1	1	1
25	Pelamar - 25	Engineering - Industri	3	3	1	1
26	Pelamar - 26	Perbankan	1	1	1	1
27	Pelamar - 27	Perhotelan	1	1	1	1
28	Pelamar - 28	Perhotelan	1	1	1	1
29	Pelamar - 29	Perhotelan	2	2	1	1
30	Pelamar - 30	Perhotelan	2	2	1	1
31	Pelamar - 31	TI (Network / Admin / Support)	1	1	1	1

32	Pelamar - 32	TI (Network / Admin / Support)	2	2	1	1
33	Pelamar - 33	Engineering - Industri	2	2	1	1
34	Pelamar - 34	TI (Network / Admin / Support)	3	3	1	1
35	Pelamar - 35	Customer Services	6	6	1	1
36	Pelamar - 36	Administrasi	13	13	1	1
37	Pelamar - 37	Sales	1	1	1	1
38	Pelamar - 38	Tukang Masak (Koki)	1	1	1	1
39	Pelamar - 39	Perhotelan	2	2	1	1
40	Pelamar - 40	Human Resources	1	1	1	1
41	Pelamar - 41	Industri Pengolahan	1	1	1	1
42	Pelamar - 42	Konstruksi	3	3	1	1
43	Pelamar - 43	Konstruksi	1	1	1	1
44	Pelamar - 44	Logistik	3	3	1	1
45	Pelamar - 45	Engineering - Industri	2	2	1	1
46	Pelamar - 46	Otomotif	1	1	1	1
47	Pelamar - 47	Otomotif	2	2	1	1
48	Pelamar - 48	Pendidikan	3	3	1	1
49	Pelamar - 49	Pendidikan	2	2	1	1
50	Pelamar - 50	Pendidikan	1	1	1	1
Rata-rata					0,94	0,94

Tabel 3.9 Nilai precision and recall sistem rekomendasi

Berdasarkan tabel 3.9 didapatkan rata-rata nilai *precision* sebesar 0,94 atau 94%. Nilai *recall* dihitung berdasarkan jumlah rekomendasi lowongan yang relevan dibagi dengan jumlah rekomendasi yang muncul. Semakin tinggi nilai *precision* maka hasil rekomendasi juga akan semakin baik. Dalam tabel tersebut terdapat 3 pelamar yang tidak dimunculkan rekomendasi oleh sistem. Hal tersebut dikarenakan data pelamar tidak match dengan data lowongan pekerjaan.

Recall digunakan sebagai ukuran rekomendasi yang relevan yang dihasilkan oleh sistem. False negative (fn) merupakan semua item relevan yang tidak dihasilkan oleh sistem. Dalam evaluasi information retrival system, recall dihitung dengan persamaan berikut [5].

4 Kesimpulan dan Saran

4.1 Kesimpulan

Kesimpulan yang didapat dari pembuatan sistem informasi rekomendasi lowongan pekerjaan ini adalah sebagai berikut:

1) Sistem berbasis web ini berhasil menampilkan lowongan pekerjaan dalam bentuk web portal
2) Sistem berhasil menampilkan rekomendasi lowongan pekerjaan berdasarkan data diri dari pelamar dan kriteria yang diinginkan oleh pelamar pekerjaan
3) Berdasarkan hasil pengujian *precision recall* sistem rekomendasi lowongan kerja menggunakan metode *euclidean distance* menghasilkan nilai *precision* sebesar 0,94 atau 94% dan nilai *recall* sebesar 0,94 atau 94%.

4.2 Saran

Adapun beberapa saran yang dapat ditambahkan dalam pengembangan aplikasi ini kedepannya adalah sebagai berikut

1) Implementasi sistem menggunakan perangkat android atau IOS
2) Dapat menambahkan menu upload persyaratan seperti KTP, CV, Sertifikat pada saat pendaftaran dan dapat dilihat oleh perusahaan penyedia kerja pada saat ada yang melamar kerja.
3) Pada bagian keterampilan dapat dibuat menjadi dinamis sehingga perusahaan dapat menambahkan kriteria kemampuan/keterampilan untuk jenis lowongan pekerjaan yang akan didaftarkan

Referensi

[1] a Henny L. (2017). *Pengembangan Website Bursa Kerja Penerapan Metode K-Nearest Neighbor Pada Penentuan Grade Dealer Sepeda Motor.* Jurnal Ilmu Pengetahuan Dan Teknologi Komputer, Vol. 2. No. 2 February 2017 E-ISSN: 2527-4864

[2] as Walpole, R. E. (1990). *Pengantar Statistika, edisi ke-3* (*Introduction to statistics*). Jakarta: PT. Gramedia Pustaka Utama.

[3] Pfleeeger, S.L. & Atlee, J.M. (2010). *Software Engineering: Theory and Practice*. 4thEdition. US:Prentice Hall.

[4] Myatt, Glenn J. 2007. Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining. Hoboken, New jersey: John Willey & Sons, Ltd.

[5] Manning, C. D., Ragahvan, P., & Schutze, H. (2009). An Introduction to Information Retrieval. Information Retrieval.

9 772301 537004

ISSN

9 772654 510006

E-ISSN