

Personality Based Lipstick Color Recommender System using K-Nearest Neighbors Algorithm

Ryan Adiputra¹, Ni Made Satvika Iswari²

Program Studi Sistem Informasi, Universitas Multimedia Nusantara, Tangerang, Indonesia

ryan.adiputra@student.umn.ac.id

satvika@umn.ac.id

Received on 18 Februari 2019

Accepted on 24 June 2019

Abstract— Lipstick is a lip color which available in many colors. A research said instant valuation of woman personality can be figured by their lipstick color choice. Therefore there is a necessity to use the right lipstick color to obtain a harmony between personality and appearance. This experiment was conducted to give lipstick color recommendation by using K-Nearest Neighbors algorithm, and Myers-Briggs Type Indicator (MBTI) personality test instrument. The system was built on Android application. Euclidean distance value is affected by 5 factors which are age, introvert, sensing, thinking, and judging. Lipstick color recommendation is obtained by fetching 7 training data with nearest Euclidean distance when compared to personality test result. The colors used in this experiment are nude, pink, red, orange, and purple. After evaluation, it is obtained the application's accuracy of 87.38% which considered as good classification, both precision and recall with 75.68% which considered as fair classification. The score for software quality is 79.13% which considered as good quality.

Keywords— K-Nearest Neighbors, Data Mining, Myers-Briggs Type Indicator, Recommender System, Lipstick.

I. INTRODUCTION

Cosmetics is needed and being used every day on every parts of human body nowadays [1]. Cosmetics has several functions for modern society which are to increase attractiveness and self confidence, protect hair and skin from the damage caused by environment, UV rays, and pollution, personal hygiene, anti aging, and to help people enjoy and appreciate life [2]. One kind of cosmetic that often being used especially by women is lipstick. Lipstick is a lip color which available in many colors. A research said instant valuation of woman personality can be figured by their lipstick color choice [3]. Therefore there is a necessity to use the right lipstick color to obtain a harmony between personality and appearance. Myers-Briggs Type Indicator (MBTI) is a personality test instrument that divides human personality based on 4 dichotomic dimensions which

are the opposite of each other where every individual has a tendency to one side of the dimensions [4]. Over the last 40 years, hundreds of MBTI applications have been implemented into various field of study with a valid and reliable result [5] and therefore is being used in this experiment. On the other hand, data mining process is done using K-Nearest Neighbors algorithm because it is the most basic and simple algorithm when the data distribution information of the relevant subject is not available [6].

II. APPROACHES

A. Myers-Briggs Type Indicator (MBTI)

MBTI is a personality test instrument which was created by Isabel Briggs Myers, and her mother, Katharine Briggs based on Carl Gustav Jung's personality theory in order to make it understandable and useful in people's lives [5]. MBTI divides human personality based on 4 dichotomic dimensions which are the opposite of each other where every individual has a tendency to one side of the dimensions which are [4] :

1. Extrovert (E) vs. Introvert (I)
2. Sensing (S) vs. Intuition (N)
3. Thinking (T) vs. Feeling (F)
4. Judging (J) vs. Perceiving (P)

From those 4 dichotomic dimensions, 16 personality combinations can be made [4]. The MBTI personality test which is used in this experiment consists of 60 questions and is divided into 15 questions for every dichotomic dimension. For each question, 2 contradictory statements will be given and each statement represents one side of the corresponding dichotomic dimension. Therefore, the value of one personality type can be obtained using the following equation [4] :

$$\text{personality} = \text{chosen corresponding statement} / 15 \quad (1)$$

B. *K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) is a classification algorithm to label an object according to its nearest distance with k amount of neighbors [7] which goal is to classify new object based on training data attributes. The steps to use this algorithm are insert training data, calculate the distances between new object and the whole training data, and determine the value of k which is the amount of nearest neighbor objects taken to classify the new object. Euclidean distance is used to calculate the distance between new object and training data [8] :

$$d = \sqrt{\sum_{i=1}^v (p_{1i} - p_{2i})^2}$$

(2)

Explanation:

- d = Euclidean Distance
- v = Amount of Dimension
- p_{1i} = Training Data Attribute Value
- p_{2i} = New Object Attribute Value

In this experiment, the Euclidean distance value is affected by 5 factors which are age, introvert, sensing, thinking, and judging. While introvert, sensing, thinking, and judging ranged from 0 to 1, age has a different scale value which are 1 if age is under or equal 20, 2 if 21 to 30, 3 if 31 to 40, 4 if 41 to 50, and 5 if above 50. Therefore, normalization is done using min-max normalization equation [9] :

$$V' = \frac{V - \min_A}{\max_A - \min_A}$$

(3)

Explanation:

- V' = New Value
- min_A = Minimum Value of a Dimension
- max_A = Maximum Value of a Dimension

The value of k is preferably an odd number [10], and can be obtained using rule of thumb equation [11]:

$$k = \sqrt{n}$$

(4)

Where n is the number of training data. This experiment used 60 training data and by using equation 4, the value of k is 7.75 while it is suggested that k is an odd number, so it turns to 7.

C. *Accuracy, Precision, and Recall*

Accuracy is the percentage of correctly classified data record after being evaluated, precision is the fraction of relevant instances among the retrieved instances, and recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances [12]. These followings are needed in order to calculate the value of accuracy, precision, and recall :

- 1. True positive (TP) = correctly identified

- 2. False positive (FP) = incorrectly identified
- 3. True negative (TN) = correctly rejected
- 4. False negative (FN) = incorrectly rejected

The calculation of accuracy, precision, and recall varies depending on how the classifications are done. In this experiment, 2 types of calculations are being used which are binary (per lipstick color calculation) and multiclass (overall calculation) classification. The formulas to calculate binary classification are as follows [12] :

Accuracy = (TP + TN) / (TP + FP + TN + FN)
(5)

Precision = TP / (TP + FP)
(6)

Recall = TP / (TP + FN)
(7)

While the formulas to calculate multiclass classification are as follows [13] :

$$\begin{aligned} \text{Accuracy} &= \frac{\sum_{i=0}^l \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{1} \\ \text{Precision} \mu &= \frac{\sum_{i=0}^l TP_i}{\sum_{i=0}^l (TP_i + FP_i)} \\ \text{Recall} \mu &= \frac{\sum_{i=0}^l TP_i}{\sum_{i=0}^l (TP_i + FN_i)} \end{aligned} \tag{9}$$

III. ANDROID APPLICATION

The Android application is developed using Android Studio, and implemented to Samsung Galaxy A7 Smartphone with Marshmallow Operating System.

A. *Main Menu*



Fig. 1. Main Menu

Main menu is used by user to navigate through the application features which are do personality test, add training data, and about.

B. Do Personality Test



Fig. 2. Do Personality Test

This Activity shows up after user pressed “Mulai Tes” Button from main menu. User has to fill all the questions which consist of age and personality test questions. Each Activity will only show 10 questions and since there are 60 personality test questions, there will be 6 pages in total. User need to press the “Next” Button to go to next page, but won’t be able to if they haven’t filled all the questions on the corresponding page. Once users reached final question, the Button will change into “Submit” which will direct user to Test Result Activity.

C. Test Result

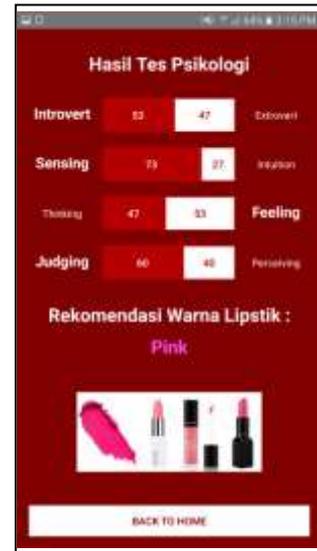


Fig. 3. Test Result

This Activity shows up after user finished the personality test and hit “Submit” Button. It shows the value from each MBTI dichotomic dimension, and giving lipstick color recommendation based on personality test result by using KNN algorithm.

D. Add Training Data

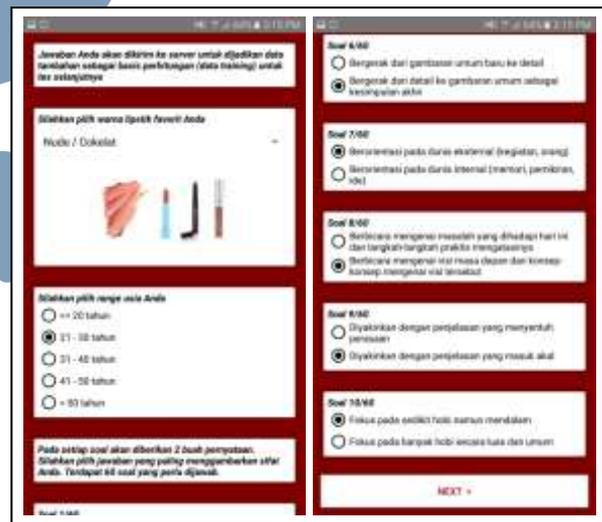


Fig. 4. Add Training Data

This Activity shows up after user pressed “Tambah Data Training” Button from main menu. This feature is used to add training data to database. Similar to Do Personality Test feature, user has to fill all the questions which consist of age and personality test questions, except with additional favorite lipstick color question. Each Activity will only show 10 questions and since there are 60 personality test questions, there will be 6 pages in total. User need to press the “Next” Button to go to next page, but won’t be able to if they haven’t filled all the questions on the corresponding page. Once users reached final

question, the Button will change into "Submit" which will send input data to database and direct user to main menu.

E. Add Training Data

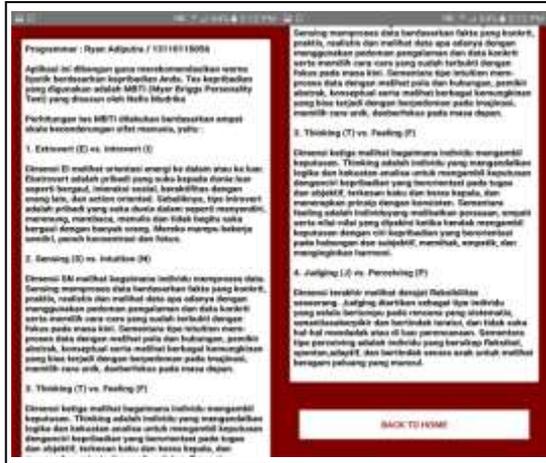


Fig. 5. About

This Activity shows up after user pressed "About" Button from main menu. It displays some information about the application as well as MBTI personality test.

IV. EXPERIMENT RESULTS

A. Implementation

The training data set used for Euclidean distance calculation was obtained from 60 respondents by filling online forms which questions include their favorite lipstick color, MBTI questionnaire, and age. The lipstick colors included in experiments were nude, pink, red, orange, and purple [14]. Out of 60, 23 respondents chose nude, 20 chose pink, 9 chose red, chose orange, and none chose purple as favorite lipstick color. Therefore, purple lipstick color cannot be discussed any further as it won't shows up as a recommendation result during testing.

On the other hand, testing was done by 37 respondents. Respondents had to do personality test using the developed Android application and put on the lipstick based on color recommendation result, then were asked if the color is suitable or not. If not, respondents were asked to put on another color until suitable color was found. Table 1 shows the test result.

TABLE I. TEST RESULT

| Id | Age | Result | Desired color | Suitable |
|-----------|------------|---------------|----------------------|-----------------|
| 1 | <= 20 | pink | pink | yes |
| 2 | 21 - 30 | nude | nude | yes |
| 3 | <= 20 | pink | pink | yes |
| 4 | <= 20 | nude | nude | yes |
| 5 | <= 20 | nude | nude | yes |
| 6 | 21 - 30 | nude | red | no |
| 7 | 21 - 30 | pink | nude | no |
| 8 | 21 - 30 | nude | pink | no |
| 9 | 21 - 30 | pink | pink | yes |
| 10 | 21 - 30 | pink | pink | yes |

| Id | Age | Result | Desired color | Suitable |
|-----------|------------|---------------|----------------------|-----------------|
| 11 | 21 - 30 | pink | pink | yes |
| 12 | 21 - 30 | nude | nude | yes |
| 13 | <= 20 | pink | pink | yes |
| 14 | 21 - 30 | pink | pink | yes |
| 15 | 21 - 30 | nude | nude | yes |
| 16 | <= 20 | nude | nude | yes |
| 17 | <= 20 | orange | orange | yes |
| 18 | <= 20 | nude | nude | yes |
| 19 | 21 - 30 | pink | pink | yes |
| 20 | 21 - 30 | nude | nude | yes |
| 21 | 21 - 30 | nude | red | no |
| 22 | <= 20 | nude | nude | yes |
| 23 | 21 - 30 | nude | nude | yes |
| 24 | 21 - 30 | pink | pink | yes |
| 25 | 21 - 30 | nude | nude | yes |
| 26 | 21 - 30 | nude | nude | yes |
| 27 | 21 - 30 | orange | red | no |
| 28 | 21 - 30 | pink | pink | yes |
| 29 | <= 20 | orange | pink | no |
| 30 | 21 - 30 | nude | nude | yes |
| 31 | 21 - 30 | pink | pink | yes |
| 32 | 21 - 30 | pink | nude | no |
| 33 | 21 - 30 | orange | nude | no |
| 34 | 21 - 30 | nude | nude | yes |
| 35 | 21 - 30 | pink | nude | no |
| 36 | 21 - 30 | pink | pink | yes |
| 37 | 21 - 30 | orange | orange | yes |

Confusion matrix can be made based on Table 1 results.

TABLE II. CONFUSION MATRIX

| | | Predicted | | | |
|--------------|--------|------------------|------|------|--------|
| | | Color | Nude | Pink | Orange |
| Color | Nude | 14 | 3 | 1 | 0 |
| | Pink | 1 | 12 | 1 | 0 |
| | Orange | 0 | 0 | 2 | 0 |
| | Red | 2 | 0 | 1 | 0 |

Summary of true positives, false positives, true negatives, and false negatives can be made based on Table 2 data.

TABLE III. SUMMARY OF CONFUSION MATRIX

| Color | True Positive | False Positive | True Negative | False Negative | Total |
|--------------|----------------------|-----------------------|----------------------|-----------------------|--------------|
| Nude | 14 | 3 | 16 | 4 | 37 |
| Pink | 12 | 3 | 20 | 2 | 37 |
| Orange | 2 | 3 | 32 | 0 | 37 |
| Red | 0 | 0 | 34 | 3 | 37 |

After obtaining Table 3 data, binary class classification accuracy, precision, and recall can be calculated using equation 5, 6, and 7.

TABLE IV. SUMMARY OF ACCURACY, PRECISION, AND RECALL BINARY CLASS CLASSIFICATION

| Warna | Accuracy | Precision | Recall |
|--------------|-----------------|------------------|---------------|
| Nude | 81.08% | 82.35% | 77.78% |
| Pink | 86.48% | 80% | 85.71% |
| Orange | 91.89% | 40% | 100% |
| Red | 91.89% | undefined | 0% |

Then, multiclass (overall) classification accuracy, precision, and recall can be calculated using equation 8, 9, and 10.

$$\begin{aligned} \text{Accuracy} &= \left(\frac{(14+16)}{37} + \frac{(12+20)}{37} + \frac{(2+32)}{37} + \frac{(0+34)}{37} \right) / 4 \\ &= 0.8783 = 87.83\% \\ \text{Precision} \mu &= \frac{(14+12+2+0)}{(14+3+12+3+2+3+0+0)} \\ &= 0.7568 = 75.68\% \\ \text{Recall} \mu &= \frac{(14+12+2+0)}{(14+4+12+2+2+0+0+3)} \\ &= 0.7568 = 75.68\% \end{aligned}$$

The application's accuracy of 87.38% can be considered as good classification, while both precision and recall with 75.68% can be considered as fair classification [12].

B. Software Quality

Software quality questionnaire was filled by 37 respondents after testing was done. The questionnaire questions were made based on ISO 9126 quality in use model. By using Pearson Product-Moment Correlation, it is obtained that all the questions are valid, and by using Alpha Cronbach formula, it is obtained the questionnaire's reliability of 0.884 which can be considered as high reliability [15]. The score for software quality is 79.13% which considered as good quality [16].

V. CONCLUSIONS & FUTURE WORKS

A. Conclusion

Personality based lipstick color recommender system using K-Nearest Neighbors algorithm has been successfully designed and developed by using MBTI personality test as its instruments. The colors used in this experiment are nude, pink, red, orange, and purple. Unfortunately, no one choose purple lipstick color upon collecting training data. Therefore purple color cannot be discussed any further. Lipstick color recommendation is obtained by fetching 7 training data with nearest Euclidean distance when compared to personality test result. After evaluation, it is obtained the application's accuracy of 87.38% which considered as good classification, both precision and recall with 75.68% which considered as fair classification. The score for software quality is 79.13% which considered as good quality.

B. Future Works

In order to improve the application, it is suggested that the application can give statistical comparison from all other users test, and also give lipstick product

recommendation based on the color recommendation result.

REFERENCES

- [1] Tranggono, R.I. and Latifah, Fatma. 2007. "Buku Pegangan Ilmu Pengetahuan Kosmetik". Jakarta: Penerbit Pustaka Utama.
- [2] Mitsui, Takeo. 1997. "New Cosmetic science". Amsterdam : Elsevier Science.
- [3] Afrisia, Rizky Sekar. 2013. "Intip Kepribadian Wanita Lewat Warna Lipstik". <http://life.viva.co.id/news/read/442053-intip-kepribadian-wanita-lewat-warna-lipstik> [Accessed 5 March 2017].
- [4] Mudrika, Nafis. 2011. "MBTI". Retrieved from : <http://docslide.net/documents/mbti-e-book-2.html> [Accessed 14 March 2017].
- [5] The Myers & Briggs Foundation. 2003. "MBTI Basics". Retrieved from : <http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/> [Accessed 14 March 2017].
- [6] Imandoust, S.B. and Bolandraftar, M., 2013. "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background". *International Journal of Engineering Research and Applications*, 3(5), pp.605-610.
- [7] Gorunescu, Florin. 2011. "Data Mining: Concepts, Models and Techniques". Berlin : Springer.
- [8] Barrett, P., 2005. "Euclidean distance: Raw, normalised, and double-scaled coefficients". Retrieved from : <http://www.pbarrett.net/techpapers/euclid.pdf> [Accessed 18 June 2017].
- [9] Moradian, M. and Baraani, A., 2009. "KNNBA: K-NEAREST-NEIGHBOR-BASED-ASSOCIATION ALGORITHM". *Journal of Theoretical & Applied Information Technology*, 6(1).
- [10] Solberg, Anne. 2013. "INF 4300 - Classification III". Retrieved from : <http://www.uio.no/studier/emner/matnat/ifi/INF4300/h13/undersvningmateriale/inf4300-2013-classification3.pdf> [Accessed 3 July 2017].
- [11] Veksler, Olga. 2010. "CS4442/9542b Artificial Intelligence II". Retrieved from : <https://www.csd.uwo.ca/courses/CS4442b/L3-ML-knn.pdf> [Accessed 3 July 2017].
- [12] Mayadewi, P. and Rosely, E., 2015. "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining". *SESINDO 2015, 2015*.
- [13] Sokolova, M. and Lapalme, G., 2009. "A systematic analysis of performance measures for classification tasks". *Information Processing & Management*, 45(4), pp.427-437.
- [14] Paramitha, Nadya. 2016. "Baca Kepribadian Dari Warna Lipstik Favorit". Retrieved from : <https://journal.sociolla.com/kepribadian-berdasarkan-warna-lipstik/> [Accessed 7 August 2017].
- [15] Janti, S., 2014. "Analisis Validitas dan Reliabilitas Dengan Skala Likert Terhadap Pengembangan SI/TI Dalam Penentuan Pengambilan Keputusan Penerapan Strategic Planning pada Industri Garmen". *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014. Yogyakarta, 15 November 2014*.
- [16] Kusuma, W.A., Noviasari, V. dan Marthasari, G.I., 2016. "Analisis Usability dalam User Experience pada Sistem KRS Online UMM menggunakan USE Questionnaire". *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 5(4).