JESICA: Journal of Enhanced Studies in Informatics and Computer Applications
Journal homepage: jesica.itsk-soepraoen.ac.id
ISSN:
Vol. 1, No. 1, February 2024, Pp. 17-22

Use Discriminant Analysis to Identify Eroticism-Related Terms in The Lyrics of Dangdut Songs

Herry Wahyu Wibowo[1], Muhammad Hasbi[2], Mochammad Anshori*[3]
1,3, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Malang, Indonesia
2, Universitas Masamus, Papua, Indonesia

***Corresponding Author**
E-mail address:
moanshori@itsk-soepraoen.ac.id

Keywords:
Identification, LDA, QDA, dangdut, machine learning

Abstract

The song "Dangdut" is one of the most popular songs in Indonesia, having gained popularity from the 1960s until the present. It's even been acknowledged as authentic Indonesian music. There are both positive and negative effects on the pendengarnya of lagu dangdut. Positive dampening can lower stress levels, and negative dampening occurs when emotions are heightened. If this was brought up by a young child who was not yet fully grown, it would give them a hard time and negatively impact their journey. According to this framework, it is recommended that any eroticism in the lyrics of dance music be identified. It is therefore advised to look for signs of sexuality in the lyrics of dangdut songs. The intention is to restrict and filter the music that kids can listen to. Using LDA and QDA classifiers in conjunction with natural language processing is the suggested approach. According to research findings, LDA can identify more than QDA. The LDA examination yielded the following results: recall = 56.522%, accuracy = 56.522%, precision = 79.13%, and F1score = 65.942%. It has been demonstrated that discriminant analysis, particularly LDA, is useful for classification, as QDA has not shown itself to be the most effective method in this instance.

1. Introduction

Songs are works of sound art that can express emotions or moods. Songs are great for self-expression since they are deeply connected to feelings, among their many other advantages [1]. However, songs can also be utilised as a form of entertainment to convey feelings that are being felt. In addition, songs can be utilised as a teaching tool in the field of education [2], [3]. Songs are engaging because they can grab the listener's interest through wordplay, lyrics, and language that speaks to societal norms. As a result, songs play a crucial part in our lives and are inextricably linked to people around us.

Songs come in a variety of forms that are also referred to as genres. Danggadut is a song genre that is commonly associated with Indonesia. One of the songs that is popular in Indonesia is dangdut. Its existence from its beginning, approximately in the year 60, to the present has been demonstrated [4]. The Malay orchestra that began to emerge in the 1970s is what gave rise to the history of dangdut songs. After that, it evolved until the early 2000s, when dangdut koplo first appeared [5]. Because the lyrics in dangdut reflected authentic Indonesian culture, the song was submitted to UNESCO in 2012 and was granted a patent as original Indonesian music [6].

Songs in the Dangdut style are distinct from other song genres in both content and lyric delivery. The benefit of dangdut music in daily life is what makes it special. Dangdut music is one of them; it can be a useful tool in therapy. It has been demonstrated that studies have indicated that listening to dangdut music can lower depression levels [7]. Rhoma Irama is one person who demonstrated how to use Dangdut songs as a da'wah medium [8].

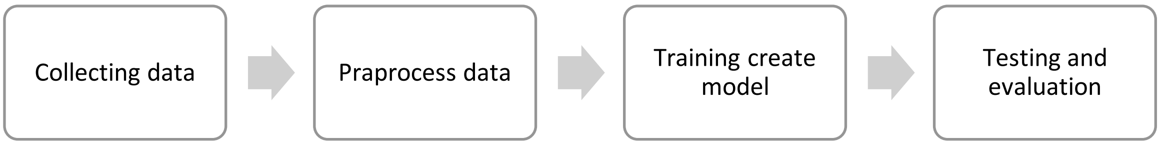
Song lyrics and dangdut music frequently enforce various values and conventions [9]. Some of the dangdut song lyrics that have recently surfaced on the Indonesian music scene, however, are less respectful—some even border on vulgar—and they emphasise erotic elements. Dangdut music used to provide a bad impression. This is so because dangdut songs are thought to have mature lyrics. There has been an incidence of sexual assault against minors as a result of the dangdut music that they were exposed to [10]. Even worse, teens' sexual behaviour may be influenced by implicit exposure to pornography [11].

Dangdut music became popular among a wide range of people, including kids and teenagers. The lyrics of Dangdut songs have both positive and negative effects. Negative dangdut song lyrics, like those containing offensive language, influence how people use language in daily life. A state of lustful arousal is called eroticism (KBBI). If young people hear lyrics containing adult and vulgar elements (eroticism), it will affect their behaviour and psychology [12]. It's feared that kids will pick up negative traits and behave badly [13]. A few more dangduts with suggestive content are likewise restricted in their distribution by the Indonesian Broadcasting Commission.

Preventive actions are therefore required. Finding and removing dangdut songs that contain eroticism is one method. There are natural language processing (NLP) techniques in computer science. NLP is frequently utilised with text data to extract the information present in it [14]. NLP can help you solve issues with sentiment analysis and classification using common language [15]. The fundamental function of natural language processing (NLP) is to preprocess text data to prepare it for machine learning (ML) classification.

Discriminant analysis, specifically linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), is the machine learning technique used in this study. The method of reducing high dimensional data through LDA is more well-known. LDA, on the other hand, can be applied to two data class classifications. LDA's ability to produce high accuracy on linearly separated data is one of its cool features [16]. While QDA and LDA are similar, QDA is used for classification on nonlinearly separated data [17]. This study will use natural language processing to handle lyric data in addition to LDA and QDA based on the previously mentioned foundation. We will compare the two methods' capabilities to determine which classifier performs the best.

2. Research Method



Collecting data

Praprocess data

Training create model

Testing and evaluation

Figure 1. Research methodology

The above Figure 1 depicts the research methodology. The initial step is gathering and labelling data based on this image. The researchers gathered lyrics to dangdut songs as their data. Some songs are not allowed to be shared or streamed because of certain lyrics [6]. This source indicates that 24 songs are only occasionally broadcast in Indonesia. The song will be categorised as erotic as a result. We will gather non-erotic songs from other dangdut songs in the interim.

Preparing the data is the second step. Lyrics from songs make up the raw data. To ensure that the data is prepared for processing using machine learning algorithms, preprocessing is required. Case folding was utilised to standardise the style of the song lyric text. Tokenization was employed to eliminate extra space, needless punctuation, and symbols. Meaningless words were filtered out. Words with affixes to their base words were returned using stemming. Finally, text data was weighted using TF-IDF to convert it into numerical form [18]. Because it helps balance the weight between words that appear frequently and words that appear rarely, the term frequency-inverse document frequency (TF-IDF) is very commonly used in weighting [19].

$$TF-IDF = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \log\left(\frac{N}{df(i)}\right) + 1 \tag{1}$$

$freq_i(d_j)$ is the number of word frequencies in the document. N is the total number of documents. $df(i)$ is the number of documents containing the word term.

Using machine learning techniques to create an identification model is the third step. Discriminant analysis is the hang method employed. Following the application of two methods, the outcomes will be contrasted. These techniques are known as quadratic and linear discriminant analysis. Equation (2) displays the LDA formula, while equation (3) displays the QDA formula. In the binary class case, the label decision is found in equation (4)[17].

$$\delta(x) \coloneqq 2(\Sigma^{-1}(\mu_2 - \mu_1))^T x + ((\mu_1 - \mu_2)^T(\Sigma^{-1}(\mu_1 - \mu_2)) + 2\, ln(\frac{\pi_2}{\pi_2}) \tag{2}$$

$$\delta(x) \coloneqq x^T(\Sigma_1 - \Sigma_2)^{-1}x + 2(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)^T x + (\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2) + \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\ln\left(\frac{\pi_2}{\pi_2}\right) \tag{3}$$

$$\hat{C}(x) = \begin{cases} 1, & if \delta(x) < 0 \\ 2, & if \delta(x) > 0 \end{cases} \tag{4}$$

The final step is testing and evaluation. There are several evaluation measures used in this research, namely accuracy to find out how good the model is at predicting, the formula is shown in equation (5); precision to find out how close the information provided by the classifier is, the formula is shown in equation (6); recall is an evaluation to determine the number of true positives predicted by the model, the formula is shown in equation (7); and f1 score is a harmonious value using calculated values from precision and recall [20]. The F1score can be applied to data that has unequal class distribution. The f1score formula is displayed in equation (8). Additionally, we compute the area under the curve, or auc, which is helpful for modelling errors from the classification model that is being developed[21].

$$Accuracy\ = \frac{TP+TN}{N}100\% \tag{5}$$
$$precision = \frac{TP}{TP+FP}100\% \tag{6}$$
$$recall = \frac{TP}{TP+FN}100\% \tag{7}$$
$$F1-score = 2 \cdot \frac{precision \times recall}{precision+recall}100\% \tag{8}$$

TP= true positive, TN = true negative, FP = false positive, and FN = false negative.

To find the classification model with the best performance, multiple evaluations are used. The evaluation value and the outcomes of the LDA and QDA will next be compared. After that, the results will be analysed and discussed.

Research design, research procedure (represented by algorithms, pseudocode, or other relevant objects), how to test, and the data collecting process are all included in the research method [22], [23]. It is necessary to identify and caption both the figure and the table. Italic type, center alignment, 10-point font size, and single spacing are used for the name and caption. Make sure that neither the name nor the caption crosses any columns or pages. Figure names and captions might go below the figure, whereas table names and captions could go above the table. The figure/table caption and name should be separated by a dot. The names of the figures and tables are given Arabic numbers, and they are then given numbers in the order in which they occur in the main text. Table 1, Tables 1, 3, and 4, Tables 3 through 5, and so forth are the proper names for the many tables. Similar names for figures should be used for Figure 1, Figure 8, and Figure 9, etc. Abbreviations should never be used in sentences that begin with a figure reference.

3. Results and Discussion

The performance of the classifier suggested in this study is used to determine the research's conclusions. Internet searches yielded the data, which was subsequently tabulated and stored. 111 instances of the data were used, and Figure 2 below displays the class distribution. It is evident from the image that up to 85 neutral data with no erotic elements exist, while 26 positive class data have eroticism.

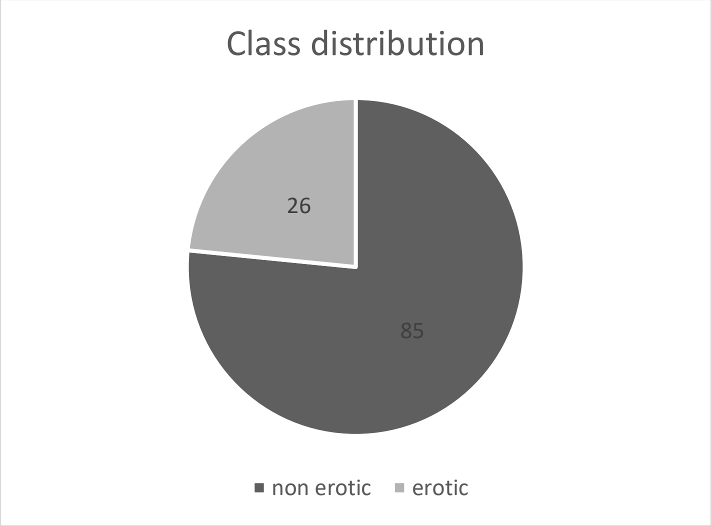


Figure 2. Class distribution of dataset

The gathered dataset consists of unprocessed data that is not yet prepared for machine learning processing. After that, case folding, tokenization, filtering, and stemming were used to preprocess the data. Once the data has made it past this point, it only has the essential lyrics from the song. Although the data is still textual, term weighting will be used to convert it to numbers this time. The TF-IDF technique is applied. Words with a high TF weight are those that show up frequently in documents. In the meantime, IDF will assign the words that appear in the document the least amount of weight [18].

Table 1. Dataset dimension

Data	Dimensional size
Before preprocessing data	111, 2
After preprocessing data	111, 1260

Table 1 above demonstrates that there were only 2 columns and 111 instances in the original data dimensions. There are currently 1260 columns in the dataset dimensions. It is evident that the dimension values have increased somewhat. Because each term will be its own column, TF-IDF generates a certain number of columns. The number of terms in the corpus determines how many columns there are as well. The dimensions will increase with the variety of terms in the corpus. The vector values obtained from the TF-IDF transformation are contained in this column.

After being vectorized, the data will be divided into 80% training and 20% test subsets. While test data is used to evaluate the model and determine its performance, training data is used to build a prediction model. Figure 3 below displays the findings of the evaluation comparison.

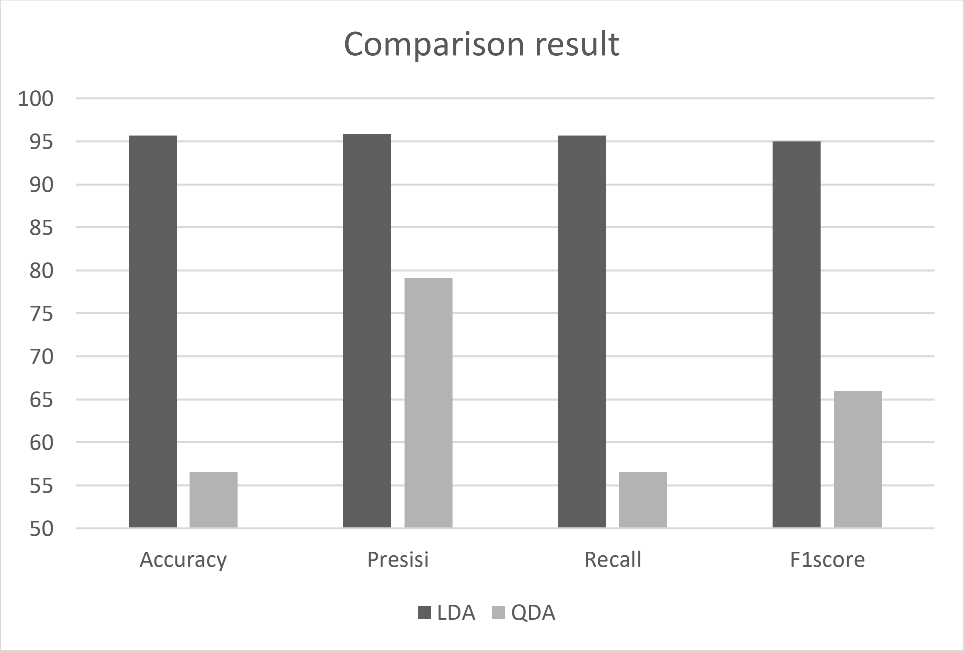


Figure 3. comparison result between LDA and QDA

The graph in Figure 3 indicates that QDA performs poorly in comparison to LDA. The results of QDA are known to be as follows: recall = 56.522%, accuracy = 56.522%, precision = 79.13%, and F1score = 65.942%. It turns out that precision yields the best results when using QDA. This demonstrates that the model has an identification error of 22.87% and a level of precision of 78.13% when it comes to identifying song lyrics with eroticism-related elements. In this instance, LDA can offer the best performance, as shown by the high evaluation value. The evaluation that is obtained is as follows: F1score = 94.978%, recall = 95.6522%, accuracy = 95.6522%, and precision = 95.85%. In this instance, it is evident from these evaluation values that LDA has the lowest prediction error/miss. The evaluation measure value, which is significantly better than QDA, demonstrates this. The data are linearly separated, as the basis above demonstrates. For the dataset problems mentioned above, LDA is therefore more appropriate and yields superior performance than QDA.

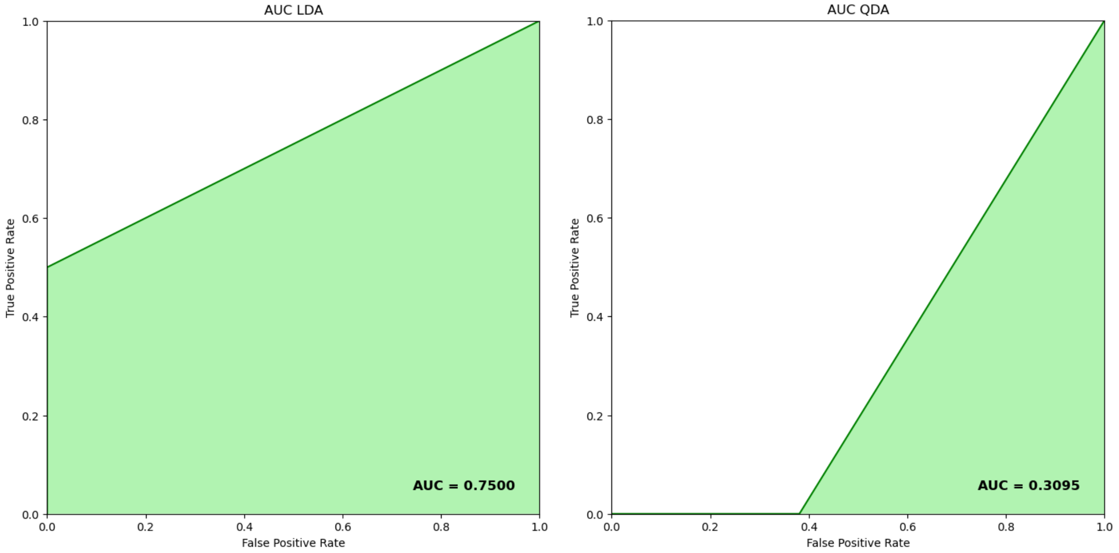


Figure 4. AUC curve of LDA & QDA

The area under the curve (AUC) between LDA and QDA is computed and displayed in Figure 4. The TPR and FPR determine the AUC number. At each thresholding, the AUC is calculated by comparing the true positive rate (TPR) to the false positive rate (FPR). According to the above figure, QDA yields an AUC value of 0.3095 and LDA yields an AUC value of 0.75. It is evident that LDA outperforms QDA in terms of prediction accuracy. Because of the weak classifier, QDA frequently produces inaccurate predictions.

4. Conclusion

In summary, we discover that erotic elements in the lyrics of dangdut songs can be recognised. Text song lyrics are preprocessed to make them mature data suitable for use as LDA and QDA training and test data. Case folding, tokenization, filtering, and stemming are a few of them. After that, TF-IDF is used to produce it as a number. Computers can only process data in numerical form, which is why this is done. The ratio used to split the data will be 80%:20%. 20% of the data is used for testing, and the remaining 80% is used as training data for LDA and QDA. Based on the test results, LDA was determined to perform the best, with the following metrics: recall = 56.522%, accuracy = 56.522%, precision = 79.13%, and F1score = 65.942%. This paper aims to investigate the effectiveness of discriminant analysis, specifically LDA and QDA, for natural language processing classification tasks. The TF-IDF vectorization results provide large dimensions, so we advise performing dimension reduction for future work.

References

[1] C. Amelia and Y. Aryaneta, “Pengaruh Musik Terhadap Emosi,” vol. 4, pp. 49–57, 2022

[2] E. Nola and D. Putri, “Integrasi Lagu dalam Rencana Pembelajaran Tematik di Sekolah Dasar,” vol. 1, pp. 53–56, 2023

[3] S. Subiyantoro and S. Mulyani, “Kegunaan Multimedia Interaktif Dalam Pembelajaran Bahasa Inggris,” *J. Edudikara*, vol. 2, no. 2, pp. 92–100, 2017

[4] D. Setiaji, “Tinjauan Karakteristik Dangdut Koplo Sebagai Perkembangan Genre Musik Dangdut,” *Handep*, vol. 1, no. 1, pp. 19–34, 2017

[5] M. F. Maulana, “Dangdut Koplo: Tubuh, Seksualitas dan Arena Kekuasaan Perempuan,” *Muqoddima J. Pemikir. dan Ris. Sosiol.*, vol. 1, no. 2, pp. 197–210, 2020, DOI: 10.47776/mjprs.001.02.07

[6] N. Vera, “Representasi Erotika Dalam Lirik Lagu Dangdut (Analisis Bahasa Kritis Terhadap Lirik LaguDangdut),” *Communication*, vol. 8, no. 1, p. 66, 2017, DOI: 10.36080/comm.v8i1.652

[7] E. Y. Lutfiani and T. Anggarawati, “Penerapan Terapi Musik Dangdut Ritme Cepat Terhadap Perbedaan Tingkat Depresi Pada Pasien Depresi Di Rsjd Dr. Amino Gondhohutomo Provinsi Jawa Tengah,” vol. 4, no. 1, pp. 16–21, 2019

[8] W. Al Basith, “Peran Dakwah Rhoma Irama Melalui Seni Musik Dangdut Tahun 1975-2003,” 2019

[9] I. Fitriyadi and G. Alam, “Globalisasi Budaya Populer Indonesia (Musik Dangdut) di Kawasan Asia Tenggara,” *Padjadjaran J. Int. Relations*, vol. 1, no. 3, p. 251, 2020, DOI: 10.24198/padjir.v1i3.26196

[10] M. H. B. Raditya, “Dangdut Koplo : Memahami Perkembangan Hingga Pelarangan,” *J. Stud. Budaya Nusant.*, vol. 1, no. 1, pp. 10–23, 2017.

[11] N. Syahruddin *et al.*, “Keterpaparan Pornografi Terhadap Perilaku Seks Remaja SMPN di Kota Tangerang Selatan,” *J. Healthc. Technol. Med.*, vol. 9, no. 1, pp. 2615–109, 2023, DOI: https://doi.org/10.24198/padjir.v1i3.26196.

[12] N. T. Rahmanda, “Perkembangan Dan Dampak Musik Dangdut Koplo Bagi Remaja Di Desa Pendowoharjo Bantul,” *Perpust. ISI Jogjakarta*, no. May, pp. 1–11, 2018.

[13] F. Gunawan, “Pornoteks dalam Lirik Lagu Dangdut: Refleksi Pendidikan Karakter Masa Kini,” *J. Ta’dib*, vol. 8, no. 1, pp. 1–18, 2015.

[14] D. R. Utari, A. Wibowo, and A. A. Sobari, “Pemrosesan Bahasa Alami pada Data Twitter untuk Penyajian Informasi Jalan dan Lalu Lintas,” *Senamika*, no. April, pp. 756–765, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/view/1419%0Ahttps://conference.upnvj.ac.id/index.php/senamika/article/downl

oad/1419/1026.

[15] B. S. Al Ar Fanny, J. M. Y. Zia Ul Haq, D. Q. Utama, and Adiwijaya, "Aplikasi Pengenalan Gejala Penyakit dengan Pemrosesan Bahasa Alami," *e-Proceeding Eng.*, vol. Vol. 8, no. No. 2, pp. 2987–2998, 2021

[16] R. Wang, "Comparison of Decision Tree, Random Forest and Linear Discriminant Analysis Models in Breast Cancer Prediction," *J. Phys. Conf. Ser.*, vol. 2386, no. 1, 2022, DOI: 10.1088/1742-6596/2386/1/012043

[17] B. Ghojogh and M. Crowley, "Linear and Quadratic Discriminant Analysis: Tutorial," no. 4, pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1906.02590

[18] Z. P. Putra and A. Nugroho, "Pebandingan Performa Naïve Bayes dan KNN pada Klasifikasi Teks Sentimen Jasa Ekspedisi," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 6, no. 3, p. 145, 2021, DOI: 10.31328/jointecs.v6i3.2635

[19] M. Hatamian, J. Serna, and K. Rannenberg, "Revealing the unrevealed : Mining smartphone users privacy perception on app markets," *Comput. Secur.*, vol. 83, no. 675730, pp. 332–353, 2019, DOI: 10.1016/j.cose.2019.02.010

[20] C. F. Suharno, M. A. Fauzi, and R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K- Nearest Neighbors Dan Chi-Square," *Syst. Inf. Syst. Informatics J.*, vol. 03, no. 01, pp. 25–32, 2017, doi: 10.29080/systemic.v3i1.191.

[21] M. Anshori, M. S. Haris, and W. Teja Kusuma, "Penerapan Backpropagation Neural Network (BPNN) Untuk Prediksi Kecanduan Smartphone Pada Remaja," *Cices*, vol. 9, no. 2, pp. 192–202, 2023, DOI: 10.33050/cices.v9i2.2701

[22] W. A. Kusuma and L. Husniah, "Skeletonization using thinning method for human motion system," in *2015 International Seminar on Intelligent Technology and Its Applications, ISITIA 2015 - Proceeding*, Aug. 2015, pp. 103–106

[23] A. E. Minarno, Y. Munarko, A. Kurniawardhani, and F. Bimantoro, "Texture Feature Extraction Using Co-Occurrence Matrices of Sub-Band Image For Batik Image Classification," in *Information and Communication Technology (ICoICT)*, 2014, pp. 249–254, DOI: https://doi.org/10.1109/ICoICT.2014.6914074