# Fine-Tuned Transformer Models for Keyword Extraction in Skincare Recommendation Systems

Ni Putu Adnya Puspita Dewi [1]*, Desy Purnami Singgih Putri [2]**, I Nyoman Prayana Trisna [3]*
*Teknologi Informasi, Universitas Udayana
adnyapuspita@unud.ac.id [1], desysinggihputri@unud.ac.id [2], prayana.trisna@unud.ac.id [3]

**Article Info**

***Article history:***

Received 2025-05-12
Revised 2025-06-17
Accepted 2025-06-19

***Keyword:***

*Keyword extraction,*
*Recommendation Systems,*
*Skincare,*
*Weigthed Jaccard,*
*nDCG.*

**ABSTRACT**

The skincare industry in Indonesia is experiencing rapid growth, with projected revenues reaching nearly 40 billion rupiah by 2024 and expected to continue to increase. The large number of products in circulation makes it difficult for consumers to find products that suit their needs. In this context, a text-based recommendation system that utilizes advances in Natural Language Processing (NLP) technology is a promising solution. This research aims to develop a skincare product recommendation system based on user needs by applying the DistilBERT model, which is specifically fine-tuned with text in the skincare recommendation domain to perform keyword extraction. The resulting keywords are then used as parameters to provide recommendations by using co-occurrence as well as using a modification of Jaccard Similarity to assess the suitability between the content and benefits of the product and user preferences. The trained extraction model achieved the best performance with a micro F1-score of 0.96 at the token level and an exact match rate of 74.25% at the entity level. The evaluation of the recommendation system showed excellent results, with an nDCG value of 0.96 and a user satisfaction rate (CSAT) of 91.9%.



## I. Introduction

Skincare is a crucial component of personal care regimens, designed to maintain skin health and address specific dermatological requirements[1]. In the context of increasing urbanization and modern lifestyles, skin condition often reflects an individual's confidence and overall health. Consequently, the ability to select the most appropriate skincare products is of significant importance. In the present day, a diverse selection of skincare products that have been clinically formulated to address specific skin concerns is available [2]. However, given the wide variety of options and the fact that every individual has different skin types and concerns, choosing the best skincare products might be difficult, this issue requires a recommendation system that may offer customized recommendations based on each individual's preferences and needs[3].

A recommendation system utilizing Natural Language Processing (NLP) offers a promising solution, as it enables the interpretation of user needs through natural language inputs, facilitating personalized product suggestions based on individual preferences.

To date, numerous studies have implemented Natural Language Processing (NLP) in skincare recommendation systems. A previous study explored the use of user input such as “Toner yang efektif menghilangkan bruntusan untuk kulit kombinasi dan berjerawat yang mengandung Centella Asiatica Extract dan PHA” as a query to generate product recommendations. In this study, the Term Frequency-Inverse Document Frequency (TF-IDF) method is applied to analyze the keywords in the query, followed by a cosine similarity calculation to calculate the similarity between the query and the products in the dataset. The results show a similarity score of 0.61 using a dataset consisting of 30 skincare product samples. This result indicates the significant potential of Natural Language Processing (NLP)-based approaches in improving the relevance of content-based recommendations, particularly in providing personalized skincare product suggestions based on users' needs and preferences [4].

Futhermore, NLP has been widely applied in recommendation systems across variety of domains, including the healthcare sector. In its implementation, NLP is used to extract features from unstructured textual data, such as user reviews, using techniques and methods such as Bag-of-Words, Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings.

Related studies indicates that NLP is effective in prescribing medications based on user-generated data, including drug attributes, demographics, and reviews, thereby demonstrating the crucial role of NLP in text-based recommendation systems [5].

In software engineering, the process of extracting information from text is one of the main focuses in the application of NLP for recommendation systems, with Named Entity Recognition (NER) being one of the employed methodologies[6]. Named Entity Recognition (NER) functions to identify and categorize predefined entities within a text and classify entities such as names of people, locations, organizations, and other entities written in text. Particularly when applied to texts from specifically trained domains, this technique exhibits a high accuracy rate.

Named Entity Recognition (NER) has shown potential in information extraction within specialized domains, especially in retrieving keywords from scientific literature. According to a study, models like DistilBERT, which have been optimized for extracting keywords from economic documents, are able to surpass the performance of conventional techniques like TextRank and TF-IDF in various evaluation scenarios. With an accuracy of 97%, a recall of 99%, and an F1-score of 98%, the domain-specific model outperforms the TF-IDF approach, which only achieves an F1-score of 42% [7]. This study reinforces that NER-based approaches and models trained in specific domains have significant advantages in keyword extraction accuracy over conventional term-based approaches.

Although NLP-based approaches for information extraction in recommendation systems have been extensively employed in various studies, there is a lack of research that applies pre-trained, fine-tuned models specifically within the skincare domain for the purpose of information extraction in recommendation systems. Consequently, this study introduces a novel methodology that utilizes domain-specific keyword extraction tailored for skincare recommendation systems. This approach is based on the observation that Named Entity Recognition (NER) and models fine-tuned on domain-specific datasets demonstrate high accuracy in keyword extraction, thereby enhancing the precision of matching user requirements with relevant skincare products.

The keywords extracted from user input are subsequently matched with skincare products based on the similarity of the active ingredients they contain. This matching process employs a similarity-based approach utilizing a modified version of the Jaccard similarity index, which quantifies the degree of similarity between the set of ingredients derived from the user input and those present in the product. A similar methodology has been applied in previous research focused on recommending skincare products, where the similarity of ingredient profiles between products was used to enhance the relevance of recommendations, yielding promising results in assisting users in identifying appropriate products [8].

## II. METHODS

This study has several steps designed to guarantee that the entire procedure proceeds in compliance with the established goals. Figure 1 shows a summary as to how the system is operating.

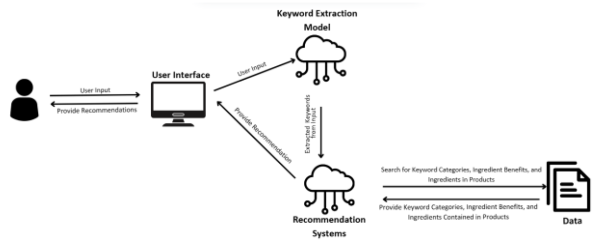


Figure 1. System Overview

### A. *Data collection*

This study utilized data from three sources: (1) web scraping results from a skincare community platform on 'X,' (2) input from 125 skincare users, each providing four distinct statements, and (3) a list of active ingredients sourced from the official Paula's Choice website, which was used to generate ingredient-based statements. The data was categorized based on four primary criteria: active ingredients, expected benefits, skin type, and product type. These criteria formed the basis for the keyword extraction labeling process. In total, 2,257 statements were collected.

### B. *Data Pre-Processing*

The data that has been collected is then processed through the preprocessing stage. The text data is initially processed through tokenization, which breaks the sentence or text into individual word units (tokens). Each token is subsequently labeled using the B-I-O (Beginning, Inside, Outside) scheme to identify tokens that form part of the keywords within a sentence. In this scheme, the label 'B' denotes the beginning of a keyword, 'I' indicates the continuation of the keyword, and 'O' signifies that the token is not part of the keyword. Keywords refer to tokens representing ingredients, benefits, product types, or skin types. An illustration of the labeling process using the B-I-O scheme is shown in Figure 2.

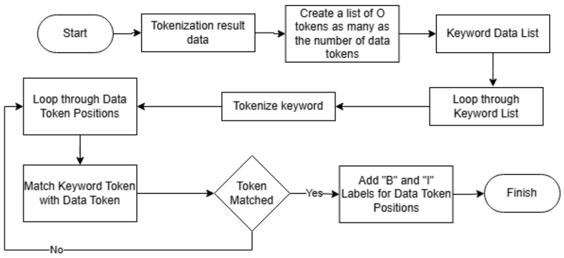


Figure 2. B-I-O Tag Process Flow

The tokens labeled using the B-I-O scheme are subsequently readjusted to maintain the validity of the labels, considering the re-tokenization process that occurs after labeling. The outcomes of this adjustment are presented in Table I. Following this, the adjusted tokens and labels are converted into numerical representations, where tokens are mapped to token IDs, and labels are converted into numerical values corresponding to the label scheme used. This step is followed by padding to ensure uniform input length across all data. The final output is a structured input comprising input IDs and attention masks, as illustrated in Table II.

TABLE I
TOKENIZATION AND B-I-O TAG ADDITION RESULTS

| Kalimat | Lower case | Tokenization | B-I-O tag |
|---|---|---|---|
| produk dengan Niacinamide dan Zinc untuk mengatasi jerawat | produk dengan niacinamide dan zinc untuk mengatasi jerawat | ['produk', 'dengan', 'niacinamide', 'dan', 'zinc', 'untuk', 'mengatasi', 'jerawat'] | [ 'O', 'O', 'B-KEYWORD', 'O', 'B-KEYWORD', 'O', 'B-KEYWORD', 'I-KEYWORD'] |

TABLE II
LABEL AND TOKEN ADJUSTMENT RESULTS

| Index | Token | Token ID | Attention Mask | Label |
|---|---|---|---|---|
| 0 | [CLS] | 101 | 1 | -100 |
| 1 | pro | 4013 | 1 | 2 |
| 2 | ##duk | 28351 | 1 | 1 |
| 3 | deng | 26957 | 1 | 2 |
| 4 | ##an | 2319 | 1 | 1 |
| 5 | ni | 9152 | 1 | 0 |
| 6 | ##ac | 6305 | 1 | 1 |
| 7 | ##ina | 3981 | 1 | 1 |
| 8 | ##mide | 24284 | 1 | 1 |
| … | … | … | … | … |
| 127 | [PAD] | 0 | 0 | -100 |

## C. *DistilBERT*

The model used in this study is DistilBERT, a distilled version of BERT (Bidirectional Encoder Representations from Transformers). DistilBERT retains about 97% of the performance of the original BERT model, with a model size that is about 40% smaller and training times that are up to 60% faster [9].

In this study, the pretrained DistilBERT model is fine-tuned to perform keyword extraction tasks based on token classification with a BIO (Beginning-Inside-Outside) labeling scheme. The dataset used has undergone preprocessing stages, including text normalization, tokenization, and manual keyword labeling before being used for model fine-tuning. The model training is conducted using the token classification architecture from the Hugging Face Transformers library, which is designed for token classification tasks based on pretrained transformer models[10].

In the fine-tuning process, the dataset was partitioned into two distinct splitting schemes: 80:20 and 70:30, with the former allocated to the training data and the latter to the testing data. Each partitioning scheme was evaluated using two distinct learning rates, namely $1\times10^{-5}$ and $2\times10^{-5}$, resulting in a total of four experimental scenarios. Furthermore, additional parameters employed during the experimentation process are detailed in Table III.

TABLE III
TRAINING PARAMETER

| PARAMETER | VALUE |
|---|---|
| Learning Rate | 1 x $10^{-5}$ dan 2 x $10^{-5}$ |
| Optimizer | AdamW |
| Batch Size | 4 Train 4 Evaluation |
| Epoch | 8 (early stopping) |
| Regularisasi | Weight decay 0,01 |

The choice of DistilBERT as the base model is driven by its computational efficiency, achieved without compromising accuracy. This has been substantiated by numerous prior studies, which have demonstrated DistilBERT's effectiveness in a variety of natural language processing (NLP) tasks, such as named entity recognition and sentiment analysis [5], [11], [12].

## D. *Model Evaluation*

This stage evaluates the model's performance in keyword extraction. The evaluation is conducted using two approaches: token-level and entity-level evaluations. At the token level, the model's performance is assessed by comparing its predictions of the label for each token in the text sequence. For each token $x_i$ in the input sequence X, the predicted label $\hat{y}_i$ is compared with the actual label $\hat{y}_i$i. From this comparison, three primary metrics precision, recall, and F1-score are calculated, considering the number of true positives, false positives, and false negatives at the token level [13].

At the entity level, precision is used to measure the proportion of relevant key phrases predicted out of all phrases identified by the model. Recall is employed to assess the model's ability to correctly identify key phrases from the ground truth. Additionally, the F1 score and accuracy are calculated to evaluate the overall percentage of correctly classified phrases out of all the predicted phrases [14].

*E. Inference Model*

The fine-tuned model is then used to extract keywords from the input sentence. The extracted keywords from the model will be used in the recommendation system. The keyword extraction process includes several steps, namely processing input sentences into tokens, converting tokens into IDs, predicting labels for each ID using the model, and mapping tokens to their corresponding labels. Furthermore, the labels are put together to form the final keyword. An illustration of the keyword extraction process is shown in Figure 3.

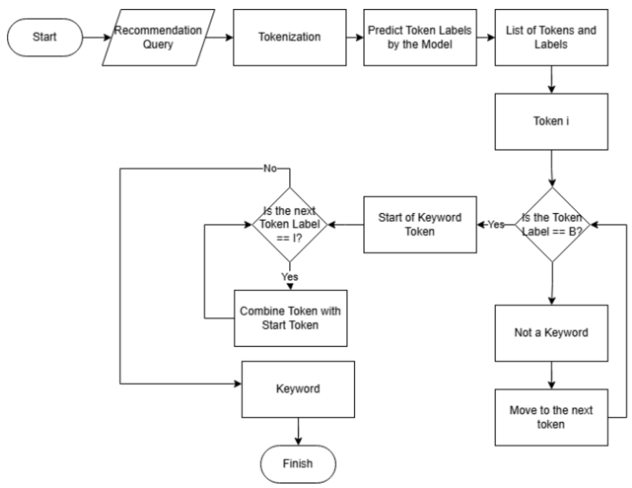


Figure 3. Inference Model

*F. Recommendation System*

The keyword extraction results are utilized as input for the recommendation system. The recommendation system utilizes the information to match skincare products based on three main aspects: active ingredients, benefits, and skin type. The matching process mainly focuses on active ingredients using the Weighted Jaccard Similarity method. The recommendation system in this research focuses on finding the product that has the highest proximity to the keywords inputted by the user. The dataset used is product data, which includes product attributes such as brand and type and also the ingredients contained, and also an ingredient dataset, which includes a list of ingredients commonly used in skincare along with the function and also the rating of each ingredient.

*1) Identifikasi Bahan Berdasarkan Co-occurrence*

If the identified keywords are skin types or benefits (e.g., “oily skin” or “moisturizing”), the system searches for the ingredients that co-occur most frequently with these keywords in the dataset. The analysis is done by calculating how often an ingredient co-occurs with a benefit or skin type keyword in the product dataset. The most frequently occurring ingredients are considered to be more relevant to the user's needs. In addition to single ingredients, the system also takes into account ingredient pairs that are often used together in products with similar benefits. This approach helps recognize ingredient combinations that are commonly found in formulations. In addition to counting occurrences, the co-occurrence results are weighted based on the rating or effectiveness of the ingredients in the dataset so that the final result considers not only the frequency of occurrence but also the quality of the recommended ingredients. The ingredient that has the highest weighted co-occurrence result will then be used to match products with keywords.

.

*2) Product Matching Using Weighted Jaccard*

Once the ingredient list is obtained, the system matches it with the composition of each product in the database using Weighted Jaccard Similarity. At this stage, the Jaccard calculation is modified by adding weights to each ingredient based on the order in which they appear in the composition list. This approach is based on the Badan POM Regulation No. 31/2018 [15], which states that the order of ingredients in the product packaging is arranged based on the largest number. Therefore, ingredients listed earlier are assumed to have higher concentrations and are given greater weight in the similarity calculation. The formula for weighted Jaccard is as follows

$$Weighted\ Similarity = \frac{\sum_{i \in I \cap P} \omega(i)}{\sum_{i \in I} \omega(i) + \ |I - P|} \quad (1)$$

- $\sum_{i \in I \cap P} \omega(i)$ is the weighted ingredient amount of the same Product as Input
- $\sum_{i \in I \cap P} \omega(i)$ is the total weight of the Product set
- $|I - P|$ is the total number of Input products that are not in the Product set

Where the weight is calculated as follows

$$\omega(i) = 1 - \frac{index(i)}{total\ amount\ of\ Product\ Ingredients} \quad (2)$$

- $i$ is the ingredient and $index$ is the position of the ingredient (starting from 0)

*G. Evaluation of recommendation systems*

The evaluation is conducted to measure the extent to which the recommendation system can provide relevant and satisfactory results based on the extracted keywords. The two metrics used are Customer Satisfaction Score (CSAT) and Normalized Discounted Cumulative Gain (nDCG).

*1) Customer Satisfaction Score (CSAT)*

CSAT is used to evaluate the level of user satisfaction with the recommendation results provided by the system. The score is calculated based on the proportion of users who give high scores to the recommendation results. The calculation formula is as follows.

$$CSAT = \frac{\sum users\ who\ gave\ a\ score\ of\ 4\ or\ 5}{Total\ user} \text{ x } 100\% \quad (3)$$

*2) Normalized Discounted Cumulative Gain (nDCG).*

The evaluation metric employed in this study is Normalized Discounted Cumulative Gain (nDCG), which is designed to assess the quality of the recommendation system by measuring the relevance of the ranked items displayed. nDCG is widely used in the evaluation of information retrieval and search systems, as it accounts for both the ranking order of the results and their relevance to the user's needs [16].

The nDCG values in this study were obtained through testing involving 125 respondents, each of whom was asked to evaluate four types of recommendation request scenarios. The four scenarios include: (1) requests based on product active ingredients, (2) requests based on product benefits, (3) requests based on skin type, and (4) free requests according to each user's preference.

In each scenario, respondents were asked to rate 5 to 10 products displayed by the system, giving relevance scores on an ordinal scale of 3 (highly relevant), 2 (relevant), 1 (less relevant), and 0 (not relevant). These scores are then used to calculate the nDCG value, which reflects how good the system is at placing the most relevant items at the top of the recommendation results.

## III. RESULT AND DISCUSSION

After fine-tuning the model using various combinations of learning rates and dataset partitions, an evaluation was conducted to assess the performance of the keyword extraction task. The evaluation employed both token-level and entity-level metrics, as presented in Table IV and Table V. Subsequently, the final model was integrated into the recommendation system, which was further evaluated to measure its overall effectiveness. The performance of the recommendation system was assessed using the Normalized Discounted Cumulative Gain (nDCG) metric, calculated based on relevance judgments provided by 125 users, comprising both male and female participants. The relevance evaluation covered a total of 500 recommendation results, considering recommendation scenarios involving 5 and 10 products, as summarized in Table VI.

### A. Token Level Evaluation

The evaluation of the fine-tuned DistilBERT model for keyword extraction using a token classification approach demonstrates a good and consistent performance across four experimental scenarios. The experiments were conducted using two data split ratios (70:30 and 80:20) and two learning rates ($1\times10^{-5}$ and $2\times10^{-5}$), with evaluation based on precision, recall, and F1-score using both micro and macro averaging.

In the first scenario (70:30 split, learning rate $1\times10^{-5}$), the model achieved optimal performance with all micro metrics reaching 0.96 and a macro F1-score of 0.94, indicating strong recognition of keywords across classes. Increasing the learning rate to $2\times10^{-5}$ in the second scenario maintained the micro metrics while slightly improving macro precision to 0.94.

TABLE IV
TOKEN-LEVEL PERFORMANCE COMPARISON ACROSS TRAINING SCENARIOS

| Dataset | Learning Rate | Metric | Accuracy |
|---|---|---|---|
| 70:30 | $1\times10^{-5}$ | Precision (Micro) | 0.96 |
| | | Recall (Micro) | 0.96 |
| | | F1-Score (Micro) | 0.96 |
| | | Precision (Macro) | 0.93 |
| | | Recall (Macro) | 0.94 |
| | | F1-Score (Macro) | 0.94 |
| | $2\times10^{-5}$. | Precision (Micro) | 0.96 |
| | | Recall (Micro) | 0.96 |
| | | F1-Score (Micro) | 0.96 |
| | | Precision (Macro) | 0.94 |
| | | Recall (Macro) | 0.94 |
| | | F1-Score (Macro) | 0.94 |
| 80:20 | $1\times10^{-5}$ | Precision (Micro) | 0.96 |
| | | Recall (Micro) | 0.96 |
| | | F1-Score (Micro) | 0.96 |
| | | Precision (Macro) | 0.93 |
| | | Recall (Macro) | 0.94 |
| | | F1-Score (Macro) | 0.94 |
| | $2\times10^{-5}$. | Precision (Micro) | 0.96 |
| | | Recall (Micro) | 0.96 |
| | | F1-Score (Micro) | 0.96 |
| | | Precision (Macro) | 0.95 |
| | | Recall (Macro) | 0.93 |
| | | F1-Score (Macro) | 0.94 |

Similarly, the third scenario (80:20 split, learning rate $1\times10^{-5}$) also showed consistent micro performance at 0.96, suggesting that the increase in training data did not affect model stability. In the fourth scenario (80:20 split, learning rate $2\times10^{-5}$), macro precision increased to 0.95, while macro recall slightly declined to 0.93. Despite this, the macro F1-score remained at 0.94, indicating balanced performance across classes.

Overall, the model consistently delivered high scores across both micro and macro evaluations. The highest macro precision was achieved in the fourth scenario, highlighting the model's ability to accurately identify domain-specific keywords. These results indicate that the fine-tuned DistilBERT model is effective in distinguishing relevant tokens for skincare recommendation contexts while minimizing the misclassification of irrelevant tokens.

### B. Entity Level Evaluation

TABLE V
ENTITY-LEVEL PERFORMANCE COMPARISON ACROSS TRAINING SCENARIOS

| Dataset | Learning Rate | Metric | Accuracy |
|---|---|---|---|
| | $1\times10^{-5}$ | Precision (Micro) | 0.78 |
| | | Recall (Micro) | 0.82 |
| | | F1-Score (Micro) | 0.80 |
| | | Precision (Macro) | 0.81 |
| | | Recall (Macro) | 0.78 |
| | | F1-Score (Macro) | 0.78 |
| | | Exact Match Rate | 73.56% |
| | $2\times10^{-5}$. | Precision (Micro) | 0.79 |
| | | Recall (Micro) | 0.83 |
| | | F1-Score (Micro) | 0.81 |
| | | Precision (Macro) | 0.83 |
| | | Recall (Macro) | 0.79 |
| | | F1-Score (Macro) | 0.79 |
| | | Exact Match Rate | 74.90% |
| | $1\times10^{-5}$ | Precision (Micro) | 0.78 |
| | | Recall (Micro) | 0.81 |
| | | F1-Score (Micro) | 0.80 |
| | | Precision (Macro) | 0.81 |
| | | Recall (Macro) | 0.78 |
| | | F1-Score (Macro) | 0.78 |
| | | Exact Match Rate | 73.49% |
| | $2\times10^{-5}$. | Precision (Micro) | 0.80 |
| | | Recall (Micro) | 0.84 |
| | | F1-Score (Micro) | 0.81 |
| | | Precision (Macro) | 0.83 |
| | | Recall (Macro) | 0.78 |
| | | F1-Score (Macro) | 0.79 |
| | | Exact Match Rate | 74.25% |

The entity-level evaluation indicates that effective tokenization plays a key role in forming accurate and coherent entity representations, with a comparison between ground truth keywords and extracted keywords based on token prediction. In this evaluation, an exact match between the correct keywords and the keywords generated by the model was used. In general, the numbers obtained show results above 0.78, with the highest exact match rate percentage reaching 74.90%, which indicates the success of the model in extracting relevant information from the text. For example, in the ground truth case of “mengurangi iritasi, astringent,” the extraction result formed by the model is “mengurangi iritasi, memberikan efek astringent” Although there is a slight difference in word arrangement, the extracted keywords, such as “reduces irritation” and “astringent,” are still relevant and acceptable in the context of the information in question. This shows that the model can effectively capture the main meaning and context in the sentence despite the structural differences in extraction.

### C. Recommendation System Evaluation

Based on the feedback provided by the respondents, the recommendation system successfully obtained a Customer Satisfaction Score (CSAT) of 91.9% for the level of user satisfaction. In addition, in more detail, users also provide an assessment of the relevance of the products generated by the recommendation system. This relevance assessment is calculated using the Normalized Discounted Cumulative Gain (nDCG) metric, the results of which can be seen in Table VI.

TABEL VI
NDCG COMPARISON ACROSS SCENARIOS

| Scena rio n | 5 Products | | 10 Products | | Total Averages |
|---|---|---|---|---|---|
| | *Woman* | *Man* | *Woman* | *Man* | |
| 1 | 0.877 | 0.927 | 0.885 | 0.930 | 0.90 |
| 2 | 0.886 | 0.922 | 0.913 | 0.933 | 0.91 |
| 3 | 0.899 | 0.897 | 0.843 | 0.896 | 0.88 |
| 4 | 0.892 | 0.910 | 0.890 | 0.966 | 0.91 |
| Total | 0.90 | | 0.91 | | |

The average relevance results presented in Table VI demonstrate the performance of the recommendation system across four different scenarios and are further analyzed based on user gender and the number of recommended products. Overall, the system exhibits strong performance, with an average nDCG value of 0.90 for recommendations involving five products, increasing to 0.91 when ten products are recommended.

Among the four scenarios, the best performance was observed in Scenario 2 (recommendations based on product benefits) and Scenario 4 (recommendations combining ingredient benefits and skin type). Both scenarios achieved an average nDCG score of 0.91, consistent across user gender and product count. In contrast, Scenario 1, which relies solely on ingredient information, yielded a slightly lower relevance score of 0.90, while Scenario 3, which focuses on skin type suitability, recorded the lowest relevance with a score of 0.88. Scenario 4 yielded the highest nDCG score due to combined information of benefits, ingredients, and skin types. However, the lower performance in scenario 3 might indicate limitations of using only skin-type data, suggesting future improvements by integrating more contextual information.

When analyzed by gender, male users consistently reported higher relevance values than female users across all scenarios. The highest relevance score, 0.96, was recorded for male users under Scenario 4 when ten products were recommended.

Regarding the number of recommended items, presenting ten product options generally led to higher perceived relevance compared to five-product recommendations, with an average nDCG of 0.91. This suggests that offering a broader set of alternatives may enhance user satisfaction with the recommendation results.

Given the characteristics of the nDCG metric, where a score of 1.0 represents an ideal ranking. The achieved average score of 0.91 indicates that approximately 91% of the recommendation ranking aligns closely with user preferences.

## IV. CONCLUSION

This research aims to develop and implement a keyword extraction methodology specifically designed for use in recommendation systems. The results indicate that the fine-tuning process of the model yields consistent outcomes, with the most optimal model identified in scenario 4, which employs an 80% training data and 20% test data split, along with a learning rate of $2 \times 10^{-5}$. This model achieved superior performance across precision, recall, and F1-score metrics, with a micro F1-score reaching 0.96. The keyword extraction technique applied within the recommendation system effectively generates relevant keywords within the context of skincare product recommendations. In conclusion, the recommendation system, utilizing keyword extraction, delivers precise recommendations based on similarity values ordered by a weighted Jaccard approach, achieving the highest nDCG evaluation result of 0.96 and a user satisfaction rate of 91.9%. However, this study presents opportunities for further refinement, particularly in integrating the results of keyword extraction into more advanced recommendation systems.

## REFERENCES

[1] K. Rodan, K. Fields, G. Majewski, and T. Falla, "Skincare Bootcamp: The Evolving Role of Skincare," *Plast Reconstr Surg Glob Open*, vol. 4, p. E1152, Apr. 2016, doi: 10.1097/GOX.0000000000001152.

[2] T. C. Lalchand and J. Joseph, "'Beakers versus botanicals' – Analyzing the efficacy of homemade skincare in comparison to manufactured skincare products," *Cosmoderma*, vol. 3, p. 164, Nov. 2023, doi: 10.25259/csdm_202_2023.

[3] E. S. Susanto, F. Hamdani, M. Anjarsari, and F. Idifitriani, "Sistem Pendukung Keputusan Pemilihan Skincare Berdasarkan Jenis Kulit Wajah Menggunakan Metode Simple Additive Weighting," *Digital Transformation Technology*, vol. 3, no. 2, pp. 786–795, Dec. 2023, doi: 10.47709/digitech.v3i2.2554.

[4] A. Sulami and V. Atina, "Penerapan Metode Content Based Filtering Dalam Sistem Rekomendasi Pemilihan Produk Skincare," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 9, pp. 172–181, Dec. 2024.

[5] S. Gousiya Begum and P. K. Sree, "Drug Recommendations Using A 'Reviews and Sentiment Analysis' By A Recurrent Neural Network," *Indonesian Journal of Multidisciplinary Science*, vol. 2, pp. 3085–3094, 2023, doi: https://doi.org/10.55324/ijoms.v2i9.530.

[6] I. K. Raharjana, D. Siahaan, and C. Fatichah, "User Stories and Natural Language Processing: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 53811–53826, 2021, doi: 10.1109/ACCESS.2021.3070606.

[7] J. Sammet and R. Krestel, "Domain-Specific Keyword Extraction using BERT." [Online]. Available: https://www.econbiz.de/Record/datenbank-econis-

[8] T. Z. L. Kyaw, S. Uttama, and P. Panwong, "Leveraging Ingredient Profiles in Content-Based Skincare Product Recommendation," in *8th International Conference on Information Technology 2024, InCIT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 319–324. doi: 10.1109/InCIT63192.2024.10810620.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[10] Hugging Fcae, "Token classification." Accessed: Apr. 19, 2025. [Online]. Available: https://huggingface.co/docs/transformers/en/tasks/token_classification

[11] S. Kofi Akpatsa *et al.*, "Online News Sentiment Classification Using DistilBERT," *Journal of Quantum Computing*, vol. 4, no. 1, pp. 1–11, 2022, doi: 10.32604/jqc.2022.026658.

[12] K. Alharbi and M. A. Haq, "Enhancing Disaster Response and Public Safety with Advanced Social Media Analytics and Natural Language Processing," *Engineering, Technology and Applied Science Research*, vol. 14, no. 3, pp. 14212–14218, Jun. 2024, doi: 10.48084/etasr.7232.

[13] W. M. Abdul *et al.*, "Named Clinical Entity Recognition Benchmark," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.05046

[14] R. Devika, S. Vairavasundaram, C. S. J. Mahenthar, V. Varadarajan, and K. Kotecha, "A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data," *IEEE Access*, vol. 9, pp. 165252–165261, 2021, doi: 10.1109/ACCESS.2021.3133651.

[15] Badan Pengawas Obat Dan Makanan Republik Indonesia, "Peraturan Badan Pengawas Obat Dan Makanan Nomor 31 Tahun 2018 Tentang Label Pangan Olahan," 2018.

[16] D. A. Nugroho, C. Lubis, and N. J. Perdana, "Film Recommendation System using Neural Collaborative Filtering Method," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 7, no. 3, pp. 926–937, Jun. 2024, doi: 10.31539/intecoms.v7i3.8033.