Data in Brief
Open access
Article template

Beauty Product reviews dataset for sentiment analysis and recommendation system

Mujahidul Islam, Nafiz Khan Tasnul, Korobi Sarkar, Noor Shat Zahan

Affiliations

East West University, Dhaka, Bangladesh

Corresponding author's email address and Twitter handle

Mimujahid6@gmail.com, www.nafizkhan.com@gmail.com, noorshatzahan99@gmail.com, nusratsarker2000@gmail.com,

Keywords

Numeric Dataset, Text Classification, Sentiment Analysis, Beauty Product Dataset, Recommender System, Natural language processing

Abstract

Product reviews help the sellers to understand their customers' expectations and sentiment towards the product and based on those reviews they take measures accordingly to heighten the satisfaction level of their customers. Beauty products are unique because various factors can influence a customer's purchase decision. With the help of machine learning techniques, the product reviews can be utilized to achieve insights and patterns to understand customers sentiment and recommend products according to their purchase records. To maintain the confidentiality of user, real dataset was not used. A synthetic dataset can heighten the efficiency of machine learning techniques. This dataset was generated by AI, packs a vast number of reviews of various products, sentiment towards those and elaborate exploratory analysis. A total of 50,000 reviews were generated from 200 different products and 1,000 unique users. A series of processing steps were performed on the raw dataset, including content addition. Another aspect of this work is that there are still not many datasets available that contains user CTR (Click-Through Rate) alongside their Spent time on a product interface which can help the researchers exploit this dataset to develop recommender systems, and natural language processing algorithms for analytical purposes.

Specification Table

Subject	Machine Learning, Data Science, Statistical Analysis and Data Mining
Specific subject area	For in-depth product performance research, this dataset which is intended for advanced machine learning tasks in the Natural Language Processing is crucial. Precise evaluations of client feelings and preferences which is essential for assessing and improving the state of current product offerings. Additionally, the

	knowledge gained from this information is invaluable for informing the creation of new goods in the future and looking into untapped markets for cosmetics.
Type of data	Categorical & numerical data
How data were acquired	In terms of research purpose, google form was created as survey. The report illustrated what every people focuses while purchasing products in an online shopping store. After carefully analysis, few attributes were selected and prepared for the data generation. These data will help building a recommender system for researchers and developers. Made the necessary calculation for different columns as per the developer's need. Then asked the AI to generate the columns based on the calculation. After the collection of data, the data was saved in excel.
Data format	Annotated, Filtered, Raw and Analyzed
Data source location	The reviews were collected at the East West University, Bangladesh, Primary Data Source- Chatgpt generated
Data accessibility	The data is publicly available.

---

1. Value of the Data:

The dataset provides user history such as user id, gender, age range, personal recommendation, CTR, spent time and an average of rating of the user. This dataset, which covers a broad spectrum of beauty items, is the first thorough compilation of product reviews. It offers distinctive perceptions into customer preferences and attitudes within the beauty sector [3]. User ratings, product names, and in-depth review explanations are painstakingly added to every review. The dataset is a valuable tool for fine-grained sentiment classification in beauty product reviews, providing precise analysis of various product features within a single review [6] and developers can use this dataset to train machine learning systems, natural language processing algorithms [5]. Tested the data set through a complete set of collaborative filtering methods and quality measures [15]. Researchers can use this dataset to create and compare automated aspect-oriented RecSys models that concentrate on particular characteristics of cosmetic products based on user preference. This aids in determining the crucial aspects of the product that influence customer pleasure or discontent [1]. This allows for comparisons with some e-commerce platforms and provides a more comprehensive visuals of consumer sentiment in various market categories [2].

2. Background

Product reviews play an important role in influencing a user's buying decision [10]. It is equally important to protect the privacy of the respective data owners [16]. This AI generated product review dataset offers a valuable resource for researchers and industry professionals to understand consumer behavior and sentiments in the rapidly evolving beauty industry including hiding the real user information. Comprehensive aspect-based hybrid (content & collaborative) recommender system and sentiment analysis are made possible by the annotations in the dataset, which provide lucid insights into particular product attributes that influence consumer happiness or dissatisfaction. This is crucial for improving product design and developing focused marketing campaigns. It also facilitates the examination of client sentiment over time, which aids in the comprehension of changes in customer preferences [6]. These days, people use a variety of channels to express their feelings, including online product reviews, which have a big impact on what consumers decide to buy [9]. Customers sharing their experiences and opinions on internet platforms has a significant impact on the growth of the beauty business. Consumers input is essential for improving products and advertising tactics. The beauty industry's growth is largely driven by online platforms where customers share their experiences and opinions, making understanding customer feedback crucial for product refinement and marketing strategies. The dataset helps organizations and researchers learn consumer preferences over time by helping with sentiment analysis, product design improvements, and focused marketing strategy development [6]. The dataset demonstrates potential for developing automated systems that can detect users' sentiments about therapies using their post reviews [5].

3. Data Description

The dataset comprises a single Excel file containing customer reviews and ratings collected based on E-commerce beauty product selling websites. This dataset includes detailed information on user reviews for various products, enabling comprehensive analysis of user behavior and product performance. After the data collection process, the dataset consists of 50,000 rows with selected attributes, meticulously annotated for detailed insights.

Table 1 illustrates the identification and usefulness of each column. This dataset includes sentences with multiple label attributes, encompassing sentiment and it has 3 sentiment classes (positive, negative, neutral) [4]. The dataset enables an in-depth understanding of customer satisfaction and product popularity. Expert annotators ensured the quality and accuracy of the sentiment labels through a thorough annotation process. In Table 2, shows the information that were analyzed from raw dataset columns such as number of total reviews of a person, number of reviews a product has, sentiments, total rating etc.

Table 1: Identification of each attribute

Column name	Column Description
Product ID	A unique identifier for each product.
Product Name	The name of the product that the review pertains to.

User ID	A unique identifier for each user who has submitted a review.
User Age Range	An age range such as 18-24, 25-34, 35-44, 45 – 54 and 55+
User Gender	User's gender whose are Male, Female and Others.
Review Title and Description	The title of the review provided by the user. The detailed review text written by the user.
Review Date and Month	Exact review date, time and month for analysis monthly sells.
Packaging Quality	Few common words for example, eco-friendly, portable, satisfying, hygienic, etc. were chosen to determine the quality of packaging.
User's Recommendation	User's overall recommendation for other users.
Product Rating	The rating given to the product by the user, typically on a scale (e.g., 1-5 stars).
Product Price	Price of the selected product.
Product Discount	A discount 5-10% based on user activity on the website
Spent Time by A User	How much time a user spent on product, checking reviews and qualities, etc.
CTR (click through rate)	It measures how often users click on a product or ad after seeing it.
Location	In which city and country, the user belongs to.
Product Type	The type of the products if it's makeup or hair treatment or skin care, etc.
Sentiment	By using shaver's model to find the sentiment of user based on their reviews.
User's Average Rating	Average of an individual user's all rating in different products
Product Average Rating	A product's average rating, rated by all user those who purchased

Table 2: Summary of the Raw Dataset

Details	Total No.
Number of Rows	**50,000**
Number of Columns	**21**
Total number of Unique products	**200**
Total number of Unique users	**1000**
Number of users who have rated more than 40 products	**917**
Number of users who have rated more than 50 products	**473**
Number of users who have rated more than 60 products	**69**

4. Experimental Design, Materials, and Methods

4.1. Dataset collection

- **Initial planning** Before generating the dataset, it was needed to validate which the dataset will perform well for developing the recommender system. According to the fig. 1, the overall process of the data generations is shown.

- **Find website and create google form:** taking the ideas from different website and choose the perfect one such as beauty product purchasing website. Create google form to know more about what things people check before purchasing a product, which reviewing system can make the justification to buy a genuine product, etc.
- **Collect Major Information and analyze the data:** After collection of forms and website observations, analyze the information and step forward to generating the dataset. Select few columns that dataset must consists, for example, User Unique ID, Name, Product Code, Review Title, Packaging quality, People spent time, Rating, etc. For a recommender system, suggesting a product to individual users it is needed that a user have previous purchase, reviews and rating histories.
- **Dataset Generating Process:** After finalizing the columns, made some prompts such as price must be in USD, should have 5-10% discount, location from different countries, if rating is low and packaging quality is also low review must be negative, etc. By utilizing the knowledge of language, structure, and data patterns acquired during training, large language models (LLMs) such as ChatGPT can create synthetic datasets. The model interprets the schema and contextually creates rows of data in response to user prompts, such as asking for a dataset with particular columns like product_id, user_id, rating, etc. This is accomplished by employing the Transformer architecture for token-by-token sampling, which keeps track of value relationships [17][18][19].
- **Dataset Finalization:** After completing the process, the dataset gets prepared for analysis. A different python program works on finding the core information about the dataset such as number of users, products, missing values, outliers, frauds etc. After the summary of the dataset is ready, it goes to a pre-processing unit for making the dataset appropriate for researchers. The extensive pre-processing of the very attribute-rich scraped data is essential to create a dataset which is sufficient for recommender systems [11]. Finally, the dataset has been saved in database for continuing further research and publication.

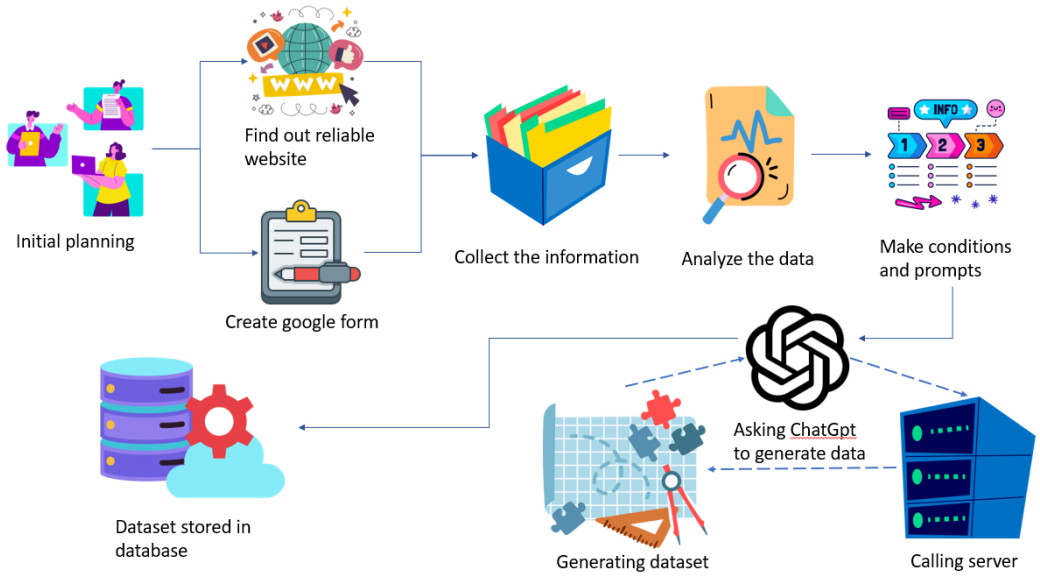

Fig-1: Data collection process

4.2. Dataset Preparation

- **Handling missing values:** To conduct our research, we began by addressing the issue of missing values in the dataset. For the missing ratings, we opted to fill these gaps using the mean values derived from the available data. This approach ensures that the overall distribution of ratings remains consistent and unbiased. When it came to missing titles, we decided to utilize the existing ratings to generate appropriate titles. For instance, a rating of 1 was assigned the title "Excellent," while a rating of 2 was labeled as "Good," and so on. Similarly, for records with missing reviews, we again used the ratings to create relevant review texts. For example, a rating of 5 was associated with the review text "I would not recommend this product," reflecting a negative sentiment.
- **Dataset Cleaning:** In preparation for data analysis, we undertook a comprehensive data cleaning process. This involved converting all text data to lowercase to maintain consistency across the dataset. Additionally, we removed any special characters that might interfere with data processing. We also trimmed any leading or trailing whitespace from text fields to ensure that the data was uniformly formatted.
- **Sentiment Analysis:** To capture further insights, the dataset was put through an emotion classification model. Considering the “User Review Title” as base the popular Shaver et al’s model [13] was utilized to annotate the reviews in 6 different emotion categories namely happy, love, no reaction, sadness, anger and fear. Finaly the outcome has divided the emotions into three different sentiment classes. First one being positive which consists of positive emotions such as, happy and love, second one expressed no emotion which is in neutral position, and the third one negative emotion such as, anger, fear and sadness [14]. The dataset is mostly comprised of reviews written in English making the dataset easier to process and clean to achieve better performance.
- **Add Necessary Attributes:** To enrich our dataset and derive more meaningful insights, we created several new attributes. We calculated the average user rating, which represents the mean rating given by each user across all their reviews. Similarly, we computed the average product rating, indicating the mean rating received by each product. We also identified the highest rating each user

had given, providing a sense of their most favorable review. It was used to explore attitudes, feelings, and moods in order to spot behavioral bias [12]. Lastly, we tallied the total number of reviews submitted by each user to gauge their overall activity.

By following these detailed steps, we were able to systematically handle missing values, derive sentiments, clean the dataset, and create valuable new attributes. In fig. 2, the pre-processing raw dataset shown visually. Finally, the comprehensive approaches not only improved the quality of our data but also enhanced our ability to analyze and interpret it effectively, leading to more accurate and insightful research findings.

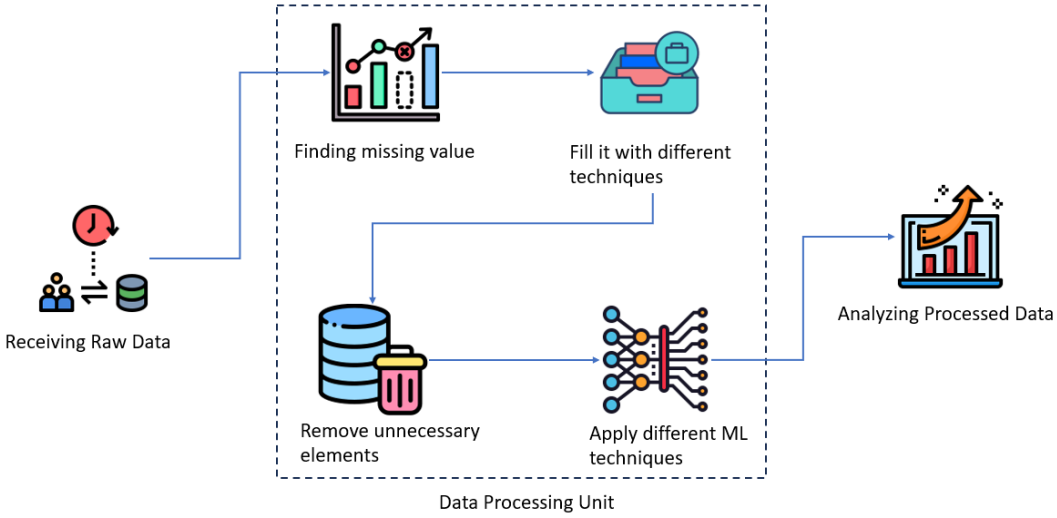

Fig-2: Pre-processing of raw dataset

4.3. Exploratory Analysis:

According to the fig. 3, the chart titled **"Top 5 Cities with Most Selling Rate"** highlights the distribution of product sales across five major global cities, based on a total of **25,333 users**. **São Paulo, Brazil** leads with **5,095 users**, making it the city with the highest sales activity. **Berlin, Germany (5,093)**, **Paris, France (5,063)**, and **Cape Town, South Africa (5,059)** follows closely, showing a relatively even distribution of sales performance across these cities. **Tokyo, Japan** has the lowest count among the group with **5,023 users**, though the difference remains marginal. Overall, the data indicates a **strong and balanced international presence**, with slight regional variations that can inform strategic decisions for targeted marketing, localized promotions, or regional inventory planning to further boost engagement and sales.

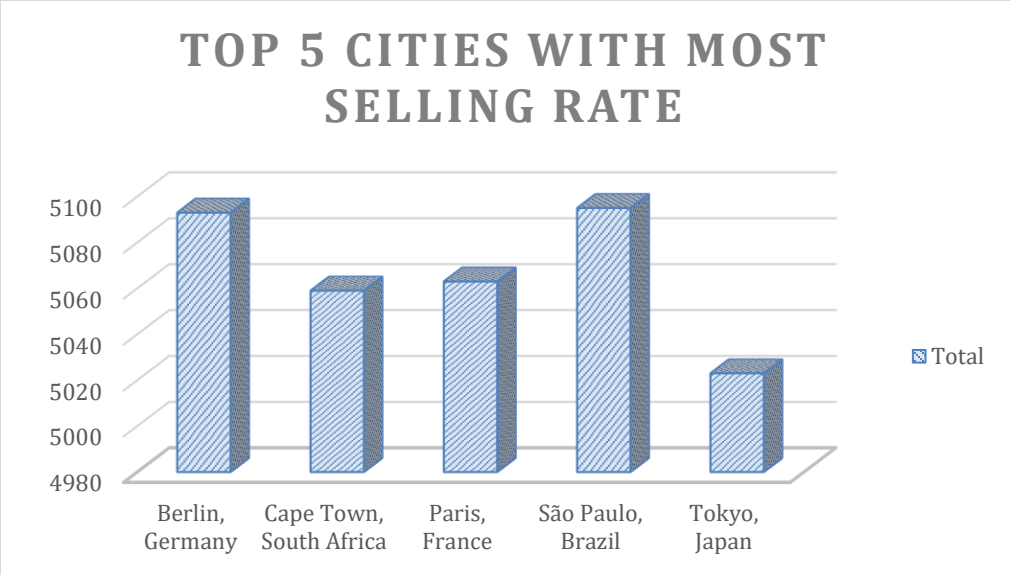

Fig-3: Distribution of product sales of top 5 cities

The pie chart titled **"Percentage of product types"** in fig.4 illustrates the distribution of 50,000 sold products across three main categories: Skin Care, Hair Treatment, and Makeup. **Skin Care products lead significantly**, accounting for **64% (31,905 units)** of total sales, making it the most dominant category. **Hair Treatment** follows with **21% (10,596 units)**, while **Makeup** holds the smallest share at **15% (7,499 units)**. This clear disparity indicates that Skin Care is the primary driver of product sales, suggesting strong customer preference and potential for continued growth in this category. Meanwhile, Hair Treatment and Makeup present opportunities for strategic development and targeted marketing to increase their market share.

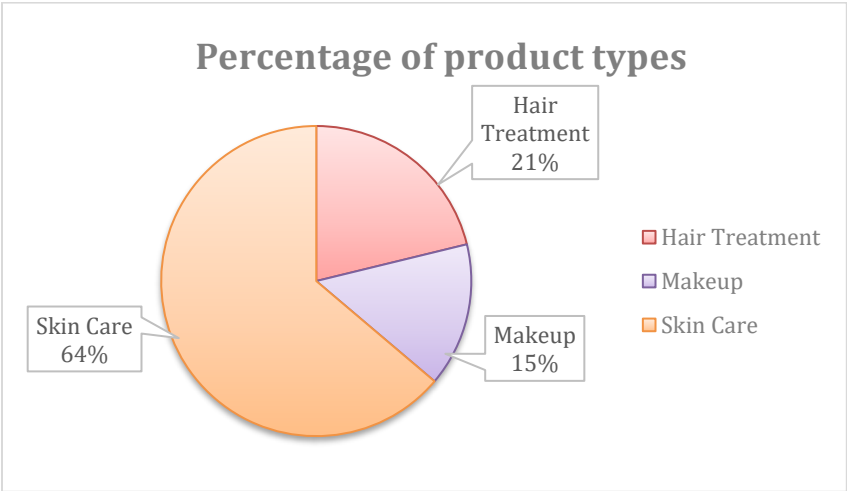

Fig-4: Pie chart of percentage of product types

The "Monthly Sale vs Income" graph reveals a clear and impactful correlation between the number of units sold and total revenue throughout the year, indicating consistent product pricing in fig-5. October emerges as the strongest month, achieving the highest sales volume and revenue, while February marks the lowest point, likely due to seasonal demand fluctuations. Spikes in revenue during May, July, and December

suggest periods of increased consumer activity, whereas June and September show moderate dips. Despite some variation, the overall trend reflects stable sales performance, with revenue closely following unit sales each month. This pattern highlights the importance of maximizing sales volume to drive income, as pricing remains steady across the year.

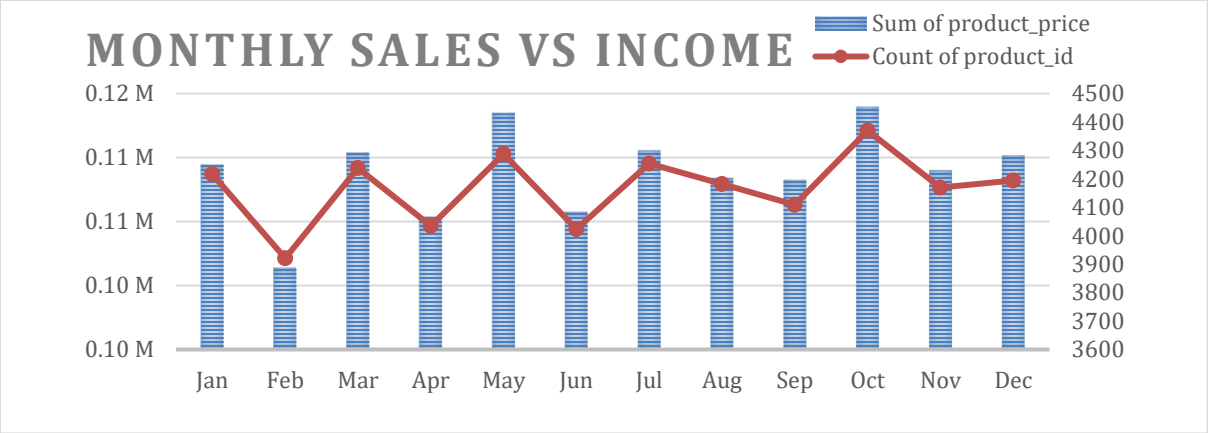

Fig-5: Graph of Monthly sales vs Income

ETHICS STATEMENT

The authors ensure that the article fulfilled:

1. The author's original work is exclusive to this publication.
2. At this point, there are no intentions to publish the essay anyplace.
3. The writers' research and analysis are presented in the article in an exact and comprehensive manner.
4. The paper appropriately acknowledges the significant contributions of the co-authors.

We certify that the aforementioned declarations adhere to Solid State Ionics' ethical standards as outlined in the statement.

CREDIT AUTHOR STATEMENT

Nafiz Khan Tasnul: System Design, data collection, & writing; Mujahidul Islam: Data collection, data processing, paper review & editing; Korobi Sarkar: Data collection, draft preparation; Noor Shat Zahan: Data collection, writing & tracking;

ACKNOWLEDGEMENTS

The authors would like to thank the owner of the site for open access.

REFERENCE

[1] Suhaimin, Mohd Suhairi Md, Mohd Hanafi Ahmad Hijazi, and Ervin Gubin Moung. "Annotated dataset for sentiment analysis and sarcasm detection: Bilingual code-mixed English-Malay social media data in the public security domain." *Data in Brief* (2024): 110663.

[2] Lee, Hyunmin, SeungYoung Oh, JinHyun Han, and Hyunggu Jung. "Creating a bias-free dataset with food delivery app reviews under data poisoning attack." *Data in Brief* (2024): 110598.

[3] Rashid, Mohammad Rifat Ahmmad, Kazi Ferdous Hasan, Rakibul Hasan, Aritra Das, Mithila Sultana, and Mahamudul Hasan. "A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews." *Data in Brief* 53 (2024): 110052.

[4] Saputra, Karen Etania. "Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review." *Data in Brief* 50 (2023): 109576.

[5] Guo, Yuting, Sudeshna Das, Sahithi Lakamana, and Abeed Sarker. "An aspect-level sentiment analysis dataset for therapies on Twitter." *Data in Brief* 50 (2023): 109618.

[6] Syed, Ayesha Ayub, Ford Lumban Gaol, Alfred Boediman, Tokuro Matsuo, and Widodo Budiharto. "A data package for abstractive opinion summarization, title generation, and rating-based sentiment prediction for airline reviews." *Data in Brief* 50 (2023): 109535.

[7] "beautylish.com" *Similarweb*, www.similarweb.com/website/beautylish.com/#overview. Accessed 24 July 2024.

[8] Sutoyo, Rhio, Said Achmad, Andry Chowanda, Esther Widhi Andangsari, and Sani M. Isa. "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks." *Data in Brief* 44 (2022): 108554.

[9] M.S. Ullal, C. Spulbar, I.T. Hawaldar, V. Popescu, R. Birau, The impact of online reviews on e-commerce sales in India: a case study, Econ. Research-Ekonomska Istraž. 34 (1) (2021) 2408–2422. (Needs proper citation)

[10] "Beautylish." *Trustpilot*, 2016, www.trustpilot.com/review/www.beautylish.com. Accessed 24 July 2024.

[11] Huda, C., Heryadi, Y. and Budiharto, W., 2024. A tourism dataset from historical transaction for recommender systems. *Data in Brief*, *52*, p.109990.

[12] Samreen, A. and Ali, S.A., 2024. Dataset construction to detect human behavior with the help of emotions, sentiments and mood for Roman Urdu. *Data in Brief*, *52*, p.109906.

[13] Shaver, Phillip, Judith Schwartz, Donald Kirson, and Cary O'connor. "Emotion knowledge: further exploration of a prototype approach." *Journal of personality and social psychology* 52, no. 6 (1987): 1061.

[14] Shaver, P.R., Murdaya, U. and Fraley, R.C., 2001. Structure of the Indonesian emotion lexicon. *Asian journal of social psychology*, *4*(3), pp.201-224.

[15] Ortega, F., Bobadilla, J., Gutiérrez, A., Hurtado, R. and Li, X., 2018. Artificial intelligence scientific documentation dataset for recommender systems. *IEEE access*, *6*, pp.48543-48555.

[16] Dandekar, A., Zen, R.A. and Bressan, S., 2018. A comparative study of synthetic dataset generation techniques. In *Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3–6, 2018, Proceedings, Part II 29* (pp. 387-395). Springer International Publishing.

[17] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, *33*, pp.1877-1901.

[18] Archana, M., Varadarajan, D.V. and Medicherla, S.S., 2022. Study on the erp implementation methodologies on sap, oracle netsuite, and microsoft dynamics 365: A review. *arXiv preprint arXiv:2205.02584*.

[19] Herrmann, F.J., Siahkoohi, A. and Rizzuti, G., 2019. Learned imaging with constraints and uncertainty quantification. *arXiv preprint arXiv:1909.06473*.