Development of a Dataset for Multimodal Fashion Recommender Models

Emmanuel A. Orisadare
Department of Computer Science and Engineering,
Obafemi Awolowo University

Idowu J. Diyaolu
Department of Family, Nutrition, and Consumer Sciences,
Obafemi Awolowo University

Iyabo O. Awoyelu
Department of Computer Science and Engineering,
Obafemi Awolowo University

ABSTRACT
Fashion recommendation systems have gained significant attention in recent years as they provide personalized and non-personalized suggestions to users based on their preferences and past behavior. The effectiveness of these systems largely depends on the availability of relevant and high-quality data, including textual, image, and other forms of data. While there are several existing datasets for fashion recommendation, they often suffer some limitations such as improper image-text mapping, small size, lack of diversity, and data quality issues. To address these limitations, this paper develops a Dataset for Multimodal Fashion Recommender Models (DMFRM-202k). The developed dataset contains an extensive collection of 202,189 fashion product images and their corresponding metadata, including product features and user ratings, preprocessed using several libraries of the Python programming language. Class labeling, feature vectors, and a ResNet50 model that was fine-tuned using transfer learning for selected fashion products are also provided. A multimodal recommender and an image classification model were developed using the DMFRM-202k dataset, the multimodal recommender model achieved an average Precision of 90% and Recall of 90% while the image classification model achieved an Accuracy of 90%, Precision of 91%, and Recall of 89% on the 10th epoch. The dataset can potentially enable researchers to develop more accurate and effective multimodal recommendation models in the fashion domain.

General Terms
Fashion Recommendation System, Digital Marketplace, Dataset, E-commerce, Machine Learning, Information Retrieval

Keywords
Multimodal, Fashion Recommender Model, Data Preprocessing, Dataset, Hybrid Recommendation, Image-based Recommendations, Convolutional Neural Network

1. INTRODUCTION
Fashion is an ever-evolving industry that constantly introduces new trends and styles. With the rise of e-commerce, consumers are increasingly turning to online platforms to purchase fashion items [1] [2]. This shift has led to the development of various fashion recommendation systems to help consumers navigate the vast array of available products. These recommendation systems use different data types, such as text and images, to suggest products to users [3]. However, from the literature review, most datasets available for developing fashion recommendation models are text- or image-based, limiting several studies for developing multimodal fashion recommendation models. Content-based recommendation models typically use textual data such as product descriptions, titles, and features. Collaborative filtering recommendation models in this domain use textual data that contains fashion product details, ratings, and users that rated the product.

On the other hand, image-based fashion recommendation models based on content-based image retrieval methods use images alone to retrieve similar fashion products. Therefore, the need for textual data is not necessary. However, developing a multimodal recommendation model that hybridizes content-based, collaborative, popularity-based, and image-based recommendation requires a dataset with images correctly mapped to respective textual data. The available literature is void of datasets with these characteristics. A few datasets with textual and image data do not include all the vital details needed to develop such a robust multimodal recommendation model. Therefore, this study has developed a dataset that will fill this gap by providing comprehensive textual and image data containing all the relevant details.

A multimodal recommendation model uses multiple modalities to provide recommendations to users. These modalities can include textual, image, audio, or other forms of data [4]. A multimodal fashion recommendation model could use product descriptions and images to provide recommendations to users. Multimodal recommendation models have several advantages over unimodal models. They can provide more accurate recommendations and better handle the cold-start problem, where there is little or no data about a new user or product [5]. This study presents the development of a dataset for a multimodal fashion recommendation model. Different modalities can be used in a multimodal fashion recommendation model, including text, images, and user behavior. Textual data such as product descriptions, titles, and features can be used to represent the attributes of fashion products. Image data can represent the visual appearance of fashion items, while user behavior data can represent user preferences and interactions with the system. Each modality has its advantages and limitations. Textual data can provide rich information about the attributes of fashion items but may not capture the visual aspects of the items [6]. Image data, on the other hand, can provide a more visual representation of fashion items but may not capture all the attributes of the items. User behavior data can provide insights into user preferences and interactions with the system but may be noisy and difficult to interpret. To overcome the limitations of individual modalities, multimodal approaches can be used, where multiple modalities are combined to generate recommendations. A recommendation model could combine textual and image data to provide more comprehensive recommendations that capture fashion items' attributes and visual appearance [7]. The importance of a good dataset in developing a multimodal fashion recommendation system cannot be overstated. A good

fashion dataset should be comprehensive, including a wide range of fashion items and user behavior data [8]. It should also be well-classified, providing detailed information about the attributes and characteristics of fashion items and user behavior data.

The contribution to the field is the development of a dataset that meets specific criteria. The dataset for the multimodal fashion recommender model (DMFRM-202k) integrates product information and user-generated content such as images and user and product details to comprehensively represent fashion products and consumer preferences. The dataset is scalable and extensible to include new data over time. This dataset provides researchers in the field with a valuable resource, and it is hoped that it will spur innovation and progress in developing multimodal fashion recommendation models. The study will contribute to digital transformation of textile and fashion design as well as sustainable production [9] [10].

2. LITERATURE REVIEW

Fashion recommendation models have been widely researched in recent years, with many studies focusing on developing models that can effectively recommend fashion items to users based on their preferences and past interactions. One of the critical challenges in developing these systems is the lack of suitable datasets, particularly those that contain both textual and image data adequately mapped. While existing datasets have been used in some studies, many contain one type of data, limiting the performance of multimodal recommendation models that require several forms of data. To address this limitation, researchers have begun exploring the development of new datasets designed explicitly for fashion recommendation. This section examined existing research on fashion recommendation systems, focusing on the aim, methodology, and respective dataset used by these studies. Additionally, the key features of a dataset appropriate for multimodal fashion recommendation will be discussed, and a review of the datasets used in related studies will be conducted.

2.1 Review of Existing Fashion Datasets

The "Fashion Product Images Dataset" [11] on Kaggle has 44,441 images of fashion products with rich metadata, including the product ID, gender, master category, subcategory, article type, base color, season, year, usage, and product display name. However, the dataset has limitations as the products lack categorization and user-generated data. The dataset can only be used for content-based recommendation models based on image similarity. Despite its limitations, it offers a diverse collection of fashion products with detailed metadata, making it useful for content-based image retrieval systems but not for collaborative or multimodal recommendation models. The "Amazon reviews on Women's dresses" [12] dataset contains over 23,000 text-based data points on women's dresses. The dataset helps develop natural language processing models for sentiment analysis or text classification, but it is limited in its suitability for fashion recommendation systems. It lacks images or URLs to download images of fashion products, making it unsuitable for multimodal recommendation systems. It also lacks user details, such as user IDs, making it challenging to build personalized recommendation systems. Lastly, it does not have columns for ratings or average ratings, making it unsuitable for developing a popularity-based recommendation system.

The "Clothing dataset (full, high resolution)" [13] available on Kaggle has 5000 images of clothing items belonging to 20 different classes with metadata containing an image, sender id, label, and kids columns. The dataset can be used to develop a content-based image retrieval model and fine-tune convolutional neural networks for better accuracy. However, the dataset is small and lacks details about the images, such as product titles or descriptions, and features, such as ratings, reviews, popularity, and user details. These limitations may hinder the development of a personalized recommendation system, and thus the dataset may not be suitable for developing a multimodal recommendation model. The "Fashion MNIST" dataset [14] on Kaggle contains 7000 grayscale fashion product images with labels from 10 classes, suitable for content-based image retrieval and machine learning for image classification. However, the dataset lacks sufficient metadata for textual description. It lacks ratings, reviews, or user information, making it unsuitable for text-based or personalized recommendation systems and multimodal recommendation models.

The "Myntra Fashion Product Dataset" [15] has 14,223 women's fashion clothing images with 11 metadata columns suitable for content-based recommendation systems based on product attributes. However, limitations include small size, focus on women's clothing, lack of user details, HTML tags, and unclassified images, making it unsuitable for building hybrid or collaborative recommendation models that incorporate user behavior and preferences. The dataset helps build content-based recommendation systems. However, developers interested in building more complex recommendation systems may need to seek additional datasets or augment this with additional data sources. The "Amazon Review Data (2018) dataset" [16] contains reviews and metadata for up to 29 categories of products on Amazon, including clothing, shoes, and jewelry. However, it has limitations, such as not including images and containing noise from scripting tags. Some image URLs are also no longer available. Additionally, the dataset requires preprocessing to be suitable for textual-based recommendation systems. It also lacks features like rating count and average rating, making it unsuitable for popularity-based recommendation systems. These limitations were addressed in the proposed dataset for the study.

2.2 Fashion Recommender Models and Dataset Used

[17] developed a fashion recommendation model that is based on visual similarity. The image processing and feature extraction are done using convolutional neural networks (CNN). The image input was initially resized and converted from RGB to BGV, then flattened, and the image's 2D matrix was converted to a vector to find the most similar images. The study reported an increase in recommendation accuracy by several percentage points compared to existing models using only visual similarity as a basis for the recommendation. The dataset used in this study contains only images of the fashion product; no textual data about the fashion products were included. Therefore, this dataset type is unsuitable for a multimodal fashion recommendation model that uses image and textual data.

[18] developed a fashion recommendation model that uses inputted images and user preferences to suggest clothing items. Users can upload an image or select favorite items from a set of clothing images saved in a preference database. The model employs CNN's AlexNet architecture for clothing type and style recognition and classification. The recommendation model provides personalized suggestions by combining the

input image and the user's preference database. The clothing recognition module of the system achieved an accuracy level of up to 91.92%. The dataset used in this study is the DeepFashion dataset. This dataset does not contain textual details of product purchases like ratings, reviews, and details of users.

[19] developed a content-based apparel recommendation model that allows users to input textual keywords such as color, type, and brand related to desired products. The model retrieves products from the dataset with the most matching textual details. The Bag of Words model, TF-IDF-based prediction, and Word2Vec models were used for text-based prediction. The dataset used in this study was obtained through web scrapping of the *amazon.com* e-commerce website. This dataset contains 183,138 data points and 19 features, including ASIN, brand, color, image URL, product type, title, price, SKU, availability, review, and images. While this dataset contains both textual and image datasets, it is only suitable for a content-based recommendation system and not a collaborative one requiring other features such as ratings, average ratings, and user details. Additionally, it was not stated in this study if the textual dataset is adequately mapped to the image dataset.

[20] developed and compared two clothing recommendation models. The first model, Collaborative Filtering, served as the baseline and calculated the closest item or user using cosine similarity. The second model used Clustering, KNN, and time series models as input for clothing recommendation. The second model achieved a higher accuracy in terms of the area under the curve (AUC) at 91%, compared to the baseline model's 85%. While the recommendation model achieved high accuracy, it can only filter using textual data. The dataset used for this model contains only textual data. The studied dataset contains more than 700,000 individual sales transactions from J. Hilburn, spanning from January 2017 to August 2019 and involving around 80,000 unique customers. The data is stored in four tables: one fact table and three dimension tables. The fact table contains all the transaction variables, including SKU, order date, customer ID, gross sales, gross units, product category, order ID, and item descriptions. The three dimension tables provide additional information about the stylists, customers, and products involved in the transactions, such as location, product price, and color. However, the dataset contains only textual data. Hence, it is not suitable for a multimodal fashion recommendation model.

[21] developed a fashion recommendation model that relies on image-based neural networks. The model employs convolutional neural networks (CNN) to suggest similar fashion items based on user image inputs. The neural network is trained using Fastai with transfer learning from ResNet50. Cosine similarity is then used to identify the dataset's most similar images. The model achieved an accuracy of up to 0.986413. The dataset used in this study was from Rent the Runway's inventory. The dataset contains some images and textual descriptions of the fashion products. However, the dataset's textual details are unsuitable for developing a collaborative recommendation model and require other details such as ratings, shoppers' info, and review details. Also, the respective image must be adequately mapped to unique IDs in the textual dataset.

3. METHODOLOGY

This section describes the methodology adopted by this study, thoroughly examining the data elicitation process and the data preprocessing approach.

3.1 Data Elicitation

In this study, developing a dataset for a multimodal fashion recommendation model involved the utilization of a parent dataset with the following characteristics. The parent dataset is a textual dataset that constitutes a subset of the extensive Amazon product data provided by [16] between 1996 and 2018. The dataset comprises 233.1 million reviews for 29 product categories, including the fashion products category labelled "clothing, Shoes, and Jewelry." The textual dataset of the "*clothing, shoes, and jewelry"* category was downloaded for this study. Two different files, namely ratings and metadata, were extracted from the dataset. These files will be the foundation for developing a dataset for multimodal fashion recommendation models. It is worth noting that using the parent dataset provides a robust foundation for the study and ensures the development of a dataset representative of the diverse range of fashion products available on the Amazon platform. The Python programming language and several libraries will be employed in developing the dataset.

The *metadata* file is a .json.gz file which contains 2,685,059 data points representing different fashion products. It contains the following features/columns: category, description, title, brand, feature, rank, date, asin, imageURL, imageURLHighRes, also_view, price, fit, also_buy, main_cat, tech1, details, similar_item, and tech2. The *metadata* dataset is depicted in Table 1.

Here is a brief description of features in the metadata dataset. *ASIN* stands for Amazon Standard Identification Number. It is a unique identifier assigned by Amazon to each product listed on its marketplace. *The title* is the name of the fashion product. The title provides a quick and easy way for customers to understand the main selling points of a product without having to read through the full product description. The *features* are typically bullet points that summarize the essential information about the product, such as its size, color, material, functionality, and any special features or benefits. The *description* provides potential customers with a clear understanding of what the product is, what it does, and how it can benefit them. *Price* is the price of the product in US dollars. *imageURL* is the URL to the product images (low resolution), while *imageURLHighRes* is the URL to the product images (high resolution). The *related* column contains related products, the fashion product. *Rank* is the rank of the product in a particular category. The *brand* features the manufacturer of the product. The *category* is a list of categories that the products belong to. A product can belong the several categories.Tech1 is the first technical detail table of the product. *Tech2* is the second technical detail table of the product. *Similar_item* is the Table of products similar to an item. *The date* is the upload date of the product. *Also_view* is a list of ASIN customers who viewed a particular product also view. *Also_buy* is a list of asin of products that customers who buy a particular product also buy. *Main_cat* is the product's main category, while details include other details of the product.

Table 1. The unprocessed metadata dataset

	Category	description	title	brand	feature	rank	asin	imageURL	imageURLHighRes
0	[clothing, shoes & Jewelry, Costumes & Aceesso …	[6” long, stretched waist measures 11 ½” acr…	Purpule Sequun Tiny Dancer Tutu Ba…	Big Dreams	[3 layers of tulle, 6” long stretched waist m…	19,963,069in Clothing, ShoesJewelry (	0000037214	NaN	NaN
1	[Clothing, Shoes & Jewelty, Luggage & Travel G…	[The Hottest Bag in Town! Brand Anello Condit…	Japen ANello Backpack Unisez L…	Anello	[Polyester Canvas waterproof Imported Size:…	4,537,420 clothing, ShoesJewelry (	0201377179	[https://images-na.ssl-images-amazon.com/image …	[https://images-na.ssl-images-amazon.com/image …

Statistics of the features/columns with their respective number of missing values are as follows: category - 0, description – 671,407, title – 50, brand – 802,651, feature – 162,958, rank – 69,814, date – 185,208, asin – 0, imageURL – 684,447, imageURLHighRes – 684,447, also_view – 1,943,585, price – 1,784,162, fit – 2,291,215, also_buy – 2,225,495, main_cat – 2,471,831, tech1 – 2,676,368, details – 2,570,076, similar_item – 2,667,932, tech2 – 2,684,995.

Total number of data per feature are as follows: category – 2,685,059, description – 2,013,652, title – 2,685,009, brand – 1,882,408, feature – 2,522,101, rank – 2,615,245, date – 2,499,851, asin – 2,685,059, imageURL – 2,000,612, imageURLHighRes – 2,000,612, also_view – 741,474, price – 900,897, fit – 393,844, also_buy – 459,564, main_cat – 213,228, tech1 – 8,691, details – 114,983, similar_item – 17,127, tech2 – 64.

3.2 Image Download and Preprocessing Technique

The metadata fashion dataset only contained textual information; hence, obtaining images mapped to the metadata for a multimodal fashion recommendation model is essential. For this purpose, a subset of the primary metadata dataset was created with only two features: *asin* and *imageURLHighRes*. Asin is the product ID, while imageURLHighRes contains a link to download the product images. The *imgURLHighRes* feature encompasses URLs that link to high-resolution images of fashion products. Since multiple images may be associated with a single product, the first URL is retained as the primary image, while the remaining URLs are discarded. Additionally, square brackets appended to the URLs are removed as part of the preprocessing pipeline. Duplicate *asins* were removed from the created data frame, and the product images were downloaded using a Python program. Some product images could not be downloaded due to issues like broken links and files not being found. Details of these missing products were removed from the dataset as they were unnecessary for the proposed model to provide valuable recommendations. 202,189 fashion product images were downloaded using this method, with each image named after the product's ASIN for easy mapping to respective metadata details. Products whose images had been downloaded (i.e, 202,189 products) were extracted from the primary metadata and rating dataset. Some features in the *metadata* dataset exhibit a substantial amount of missing values. Most of these features are deemed unimportant to developing a multimodal fashion recommendation model.

Moreover, their integration into the model could potentially undermine the accuracy of recommendations. Consequently, these features were removed from the dataset. The features in question are as follows: *also_view, date, price, fit, also_buy, main_cat, tech1, details, similar_item, and tech2*. The new dataset contains 202,189 rows with 9 columns: *category, description, title, brand, feature, rank, asin, imageURL, and imageURLHighRes.* The *imageURL* feature, which contains links to download low-resolution images of fashion products, is insufficient for our needs. Consequently, the *imageURL* was removed from the dataset. The new shape of the dataset is as follows: 202,189 rows and 8 columns. In order to make the textual features of the dataset suitable for usage by recommendation algorithms, there is a need to preprocess these features using several techniques. Data preprocessing activities performed on the features of the textual dataset include decapitalization, removal of stop words and punctuations, stemming, and lemmatization.

Furthermore, certain activities are tailored to specific features during the data preprocessing stage, as outlined below. The *feature* column initially contained square brackets enclosing the product feature, which were removed. Similarly, the description and category fields included square brackets that were subsequently removed. Each product in the dataset is assigned a main category, namely *clothing, shoes, and jewelry*, alongside subcategories. As the main category appears in every product, a clothing product, for instance, may be associated with clothing, shoes, and jewelry. This implies that the "shoes" and "jewelry" subcategories could impact the recommendation process. Consequently, the string "clothing, shoes, & jewelry" has been eliminated, and the products have been remapped to their corresponding subcategories. For the 'rank' feature, the value may include the string "570inClothing,ShoesJewelry," which needs to be removed to ensure the field can be used as an integer for faceted price classification in recommendation models. Furthermore, Feature engineering was performed on the dataset by combining several key features, including category, description, feature, title, and brand, to create a new feature called "*imp_features*." This newly created column serves as a collection of essential attributes integral to the content-based recommendation system. The new shape of the metadata file is as follows: 202,189 rows and 9 columns.

The rating dataset comprises four features: a*sin, user_id, rating, and timestamp*. The *asin* serves as a unique identifier for the product, which matches the asin value in the metadata dataset. Similarly, *user_id* is a distinct identifier for each user on Amazon. The *rating* feature denotes a user's numerical score (from 1 to 5) on a particular product. Lastly, the timestamp feature specifies the date and time at which the user rated the product. At the time of download, the rating dataset was comprised of 32,292,099 rows with four columns. The rating dataset is represented in Table 2.

Table 2. The rating dataset

	asin	user_id	rating	timestamp
0	0871167042	A3OT9BYASDGU2X	4.0	1398470400
1	0871167042	A28GK1G2KDXHRP	5.0	1397692800
2	0871167042	A3NFXFEKW8OK0E	5.0	1397606400

The timestamp feature was considered unimportant in developing a multimodal fashion recommendation model. Hence it was removed. After removing the timestamp feature, the ratings dataset's new shape is 32,292,098 rows and 3 columns. From the dataset, it can also be observed that different users can rate one product, and users can rate multiple products they have purchased. Therefore, there is a need to calculate the number of ratings for each unique asin in the rating dataset. Also, there is a need to get an average rating of individual products. The result of this using Python is represented in Table 3.

Table 3. Products with rating count and an average rating

	asin	user_id	rating	rating_count	avg_rating
0	0871167042	A3OT9BYASDGU2X	4.0	39	4.564103
1	0871167042	A28GK1G2KDXHRP	5.0	39	4.564103
2	0871167042	A3NFXFEKW8OK0E	5.0	39	4.564103

From Table 3, it is evident that multiple users can rate a single product. Nevertheless, to obtain a dataset that consists of distinct asin, rating counts, and average ratings, the groupby function of pandas was utilized. The outcome of this process is represented in the following Table (Table 4)

Table 4. Products with respective rating counts and an average rating

asin	rating_count	avg_rating
0000037214	1	1.0000
0201377179	4	4.0000

Figure 1 shows the rating ranges (1 to 5) distribution for products in the rating dataset.

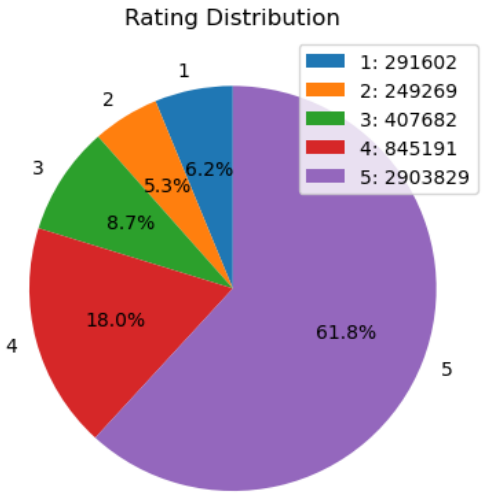

Figure 1: Rating Distribution

Given that only 202,189 products whose images (in JPG format) have been downloaded are in the metadata dataset, it only makes sense to extract only the ratings of these products from the entire rating dataset. This leaves the rating dataset with 4,697,573 data points with 3,117,073 distinct users and 202,189 different products. Finally, the metadata dataframe was merged with the latest gotten rating dataframe, which contains the unique product asin (id) with their respective rating_count, and avg_rating.

3.3 Classification of the Fashion Product Images

50,000 of the over 202,189 downloaded images have been classified into major categories. The distribution of these images into different classes is as follows: clothing: 6,674; footwear: 23,154; jewelry: 5,969; watches: 1,556; caps: 778; glasses: 361; belts: 124; bags: 2,561; ties: 184; wallets: 520; socks: 27; gloves: 99; and unclassified: 7993. The clothing contains images of clothing of males and females. The types of clothing present here include T-shirts, dresses, jeans, jackets, skirts, sweaters, suits, and shorts. The footwear contains different types of footwear for adults and children. These include sneakers, boots, sandals, high heels, loafers, flip-flops, running shoes, and slippers. Jewelry contains images of jewelry for males and females. These include necklaces, earrings, bracelets, rings, brooches, anklets, watches made of precious metals, and cufflinks.

Watches contain images of wristwatches and accessories. Caps contain different types of caps, such as baseball caps, beanies, bucket hats, flat caps, snapback caps, trucker hats, dad hats, cowboy caps, and other designer caps. Glasses contain images of different types of fashion, recommended and sunglasses. Belt contains images of belt for clothing such as dress belts, casual belts, webbing bets, stretch belts, and so on. Bags contain different types of bags such as backpacks, messenger bags, satchels, clutches, cross body bags, duffel bags, traveling bags, and other fashion bags. Ties contain images of different types of types such as neckties and bowties. Wallet contains images of different type of wallets such as bi-fold wallet, tri-fold wallet, card holder wallet, travel wallet, wrist wallet, coin wallet, and phone wallet. socks contain images of different types of socks such as ankle socks, crew socks, knee-high socks, dress socks, athletic socks, toe socks, and no-show socks. Gloves contain images of work, winter, driving, gardening, and fashion gloves. Unclassified – This class of

images were tagged unclassified because visually, they do not belong to anything that can be labelled as a fashion product. Examples of these include signposts, images containing textual descriptions, masks, umbrella, bicycles, and product boxes.

4. THE DMFRM-202k DATASET

This section describes the preprocessing output of the DMFRM-202k dataset, a collection of linked textual and image data suitable for various recommendation models, including multimodal ones. The dataset includes several sub-datasets, each described as follows. The *Metadata1* dataset is a CSV file that contains unprocessed textual details of 202,189 fashion products with the following columns: category, description, title, brand, feature, rank, asin, imageURLHighRes, and imp_features. *Metadata2* dataset is a CSV file that contains all the columns of *Metadata1*; however, decapitalization, lemmatization, stemming, removal of stop words, and punctuation have been applied to these columns. *Ratings1* is a dataset comprising 4,697,573 ratings of 202,189 products from 3,117,073 users, with columns asin, user id, and ratings. The *Rating1* dataset has an average user interaction count of 1.5 and average product interaction count of 23.2. *Ratings2* is a CSV file that includes ratings1 and additional columns for rating count and the average rating for each product. *Metadata_ratings* is a CSV file that combines *Metadata2* with each product's rating count and average rating. The *Metadata_ratings* linked to respective images through the *asin* is represented in Table 5.

Table 5. Metadata merged with rating count and an average rating

	category	description	title	brand	feature	rank	asin	imgURLHighRes	raing _count	avg _rating	imp_features
0	cloth shoe jewelri luggag travel gear backpack…	the hottest bag in town brand anello condit 10…	japan anello backpack unisex larg light gray b…	anello	polyest canva waterproof import size larg h 40…	4537420	021377179	https://images-na.ssl-images-amazon.com/image…	4	4.0	japan anello backpack uniex larg light gray b…
1	cloth shoe jewelri luggag travel gear backpack	the hottest bag in town brand anello condit 10…	japan anello backpack unisex beig larg pu…	anello	Pu leather import size larg pu leather h 39 x …	3994472	0204444454	https://images-na.ssl-images-amazon.com/image…	4	4.5	japan anello backpack unisex pink beig larg pu…
2	cloth shoe jewelri luggag travel gear backpack…	the hottest bag in town brand anello condit 10…	japan anello backpack unisex black larg pu lea…	anello	Pu leather import size larg pu leather h 39 x …	635761	0204444403	https://images-na.ssl-images-amazon.com/image…	9	5.0	japan anello backpack unisex black larg pu lea…

The All_fashion_product_images ZIP file contains 202,189 unclassified images of fashion products. The *Feature_vector_pkl* file is a pickle file that contains feature vectors extracted using the ResNet50 model of the CNN for the first 100,000 images in the All_fashion_product_images file. The filename_pkl is a PKL file that contains the respective file names of the feature vectors extracted in the Feature_vector_pkl file. The *13_classes_dataset* is a ZIP file that contains images of the first 50,000 images in the *All_fashion_product_images* file categorized into 13 classes. The *Train_test_validation* ZIP file contains 22,000 images from 4 classes (footwear, clothing, jewelries, and bags) split into train, test, and validation datasets in the ratio 80:10:10. Due to class imbalance; data augmentation was applied to the bags class to increase the number of images to 5,500. At the same time, downsampling was performed on the footwear, jewelries, and clothing classes to reduce their respective images to 5500. Additionally, data augmentation was applied to the training set only. The DMFRM-202k dataset is available publicly and can be downloaded from [22] under the Creative Common License. The graphical distribution of the 50,000 images into 13 different classes is given in Figure 2.

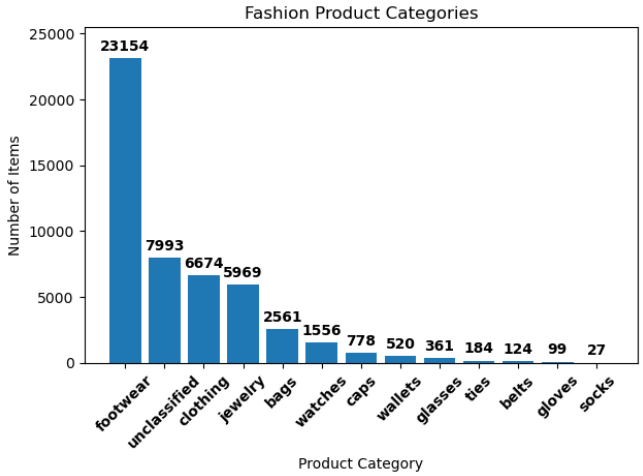

Figure 2: Distribution of the Fashion Product Images

A comparison of the DMFRM-202k dataset with the existing dataset for fashion recommendation models is given in Table 6.

Table 6: DMFRM-202k dataset versus existing fashion datasets (MD – modality, M – Mapped?, P- number of products, U – number of users, R – number of ratings, I – number of images, Img – image, Txt – textual, N/A – not available)

Dataset	MD	M	P	U	R	I
Myntra Fashion Product Dataset [18]	Img, Txt	No	14,331	N/A	N/A	14,500
Fashion Product Images Dataset [13]	Img, Txt	Yes	44,446	0	0	44,441
Amazon reviews on Women's dresses [14]	Txt	No	23,487	N/A	23,487	0
Clothing dataset (full, high resolution)" [15]	Img, Txt	Yes	5,404	0	0	5,762
Fashion MNIST [16]	Img, Txt	No	70,000	0	0	70,000
Amazon Review Data (2018) [17]	Txt	No	1,998,492	12,483,678	31,697,963	0
DMFRM-202k	**Img, Txt**	**Yes**	**202,189**	**3,117,073**	**4,697,573**	**202,189**

5. EXPERIMENT

A multimodal recommender model was developed using the DMFRM-202k dataset. The model uses image and textual data as a query to provide recommendations. Firstly, the model takes the input image and finds K of the most visually similar images using the fine-tuned CNN ResNet50 model, the Nearest-neighbor algorithm, and the Euclidean distance similarity metric. Secondly, these K images' textual details (imp_features), which merge category, title, feature, and description, are extracted and preprocessed. The model then takes the textual query of the user, preprocesses it, and string matches it with the K visually similar images using the RapidFuzz library in Python. The most matching product is selected as the first multimodal recommendation. Using the TF-IDF vectorizer, the model vectorizes the imp_feature of this most matching product and the other K-1 most visually similar images. The similarity score between the vector representation of the most matching product from RapidFuzz and the other K-1 visually similar images is calculated and sorted using Cosine similarity and output as recommendations to the user. Different images and textual input queries were combined as a query to test the model, and the model achieved an average Precision and Recall of 90% and 90% respectively.

An image classification model was also developed using images from the DMFRM-202k dataset. The task involved fine-tuning a ResNet50 model using transfer learning on a dataset of fashion product images. The original ResNet50 model had been pre-trained on a large dataset like ImageNet and was adapted for the specific fashion product classification task. The dataset consisted of four of the initial 13 classes: bags, clothing, footwear, and jewelry. The training set contained 4400 images per class to ensure equal representation, while the validation and test sets comprised 550 images per class. Data augmentation and downsampling techniques were employed to balance the number of images across classes. Starting with the pre-trained ResNet50 model, the output layer was modified to accommodate the four classes in the fashion product classification task. This involved adjusting the output layer to match the desired number of classes. In order to preserve the valuable knowledge learned by the pre-trained model, the initial layers (convolutional layers) were frozen, and only the newly added layers (including the output layer and possibly additional layers) were trained during the fine-tuning process. The model was trained using the fashion product dataset, feeding the training images through the network, calculating the loss (cross-entropy loss in this case), and updating the weights of the trainable layers through backpropagation. Multiple epochs of training were performed. After training, the performance of the fine-tuned model was evaluated on the validation and test set. On the 10th epoch, the model achieved an accuracy of 0.90 (90%), precision of 0.91 (91%), and recall of 0.89 (89%). These metrics provide insights into the model's ability to classify fashion product images.

Other recommendation models that can be developed using the DMFRM-202k dataset include collaborative, popularity-based, content-based, hybrid, and multimodal recommendation models. Using transfer learning, the developed dataset can be used to fine-tune existing CNN models such as ResNet50, InceptionV3, VGG16, and VGG19. This will help improve classification accuracy and make the fine-tuned model well-suited for the desired usage.

6. CONCLUSION, LIMITATIONS, AND RECOMMENDATIONS

The result of this study is a DMFRM-202k dataset consisting of 202,189 images with respective metadata and 4,697,573 ratings from 3,117,073 users. It also includes a fine-tuned ResNet50 model and other sub-datasets. The primary scope of this dataset is to support the development of multimodal fashion recommendation models. This is probably the first large-scale dataset in the fashion recommendation system community that provides accurately mapped textual and image datasets along with other features such as ratings, image classification, feature vectors, and dataset split into the train, validation, and test sets. The developed dataset is rich and helpful in developing various recommendation models. However, some limitations must be acknowledged. The textual data lacks user demographic information, such as age, gender, and location. This absence of demographic data may limit the developed models' effectiveness in making demographic recommendations relevant to specific user groups. It is recommended that future studies develop datasets that will include demographic information about users in order to improve the accuracy of the models' recommendations. Recommendations for future study include further classifying the main classes of fashion products into subclasses. This will improve the granularity of the dataset and may lead to more accurate recommendations for users with specific preferences for certain types of fashion products. Additionally, future studies should focus on downloading more images of the fashion products in the dataset to ensure that multimodal recommendation models can access complete data. Overall, these recommendations will help address the current dataset's limitations and lead to the development of more effective multimodal recommendation models for the fashion industry.

7. REFERENCES

[1] Sumarliah, E., Usmanova, K., Mousa, K. and Indriya, I. 2022. "E-commerce in the fashion business: the roles of the COVID-19 situational factors, hedonic and utilitarian motives on consumers' intention to purchase online," International Journal of Fashion Design, Technology and Education, vol. 15, no. 2, pp. 167-177

[2] Diyaolu, I. J., Obayomi, E. O., and Bamidele, T. A. 2019. Influence of Information and Communication Technology on Dress Culture among Senior Secondary School Students in Osun State, Nigeria. Nigerian Journal of Textiles, 5, 37-41

[3] Chakraborty, S., Hoque, M. S., Rahman Jeem, N., M. C.Biswas, M. C., Bardhan, D. and Lobaton, E. 2021. "Fashion recommendation systems, models and methods: A review" Informatics, vol. 8, no. 3, p. 49.

[4] Zhou, X., 2023"MMRec: Simplifying Multimodal Recommendation," *arXiv* preprint arXiv:2302.03497,

[5] Baig, M. Z. and Kavakli, M. 2020."Multimodal systems: taxonomy, methods, and challenges," *arXiv* preprint arXiv:2006.03813.

[6] Al-Halah, Z,. Stiefelhagen, R. and Grauman, K. 2017. "Fashion forward: Forecasting visual style in fashion," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 388-397.

[7] Wu, Q., Zhao, P., and Cui, Z. 2020. "Visual and textual jointly enhanced interpretable fashion recommendation," IEEE Access, vol. 8, pp. 68736-68746.

[8] Zou, X. Kong, X. Wong, W. Wang, C., Liu, Y. and Cao, Y. 2022. "Fashionai: A hierarchical dataset for fashion understanding," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0-0.

[9] Ogunyemi, A. A., Diyaolu, I. J., Awoyelu, I. O., Bakare, K. O., Oluwatope, A. O. 2023. Digital Transformation of the Textile and Fashion Design Industry in the Global South: A Scoping Review. In: Saeed, R. A., Bakari, A. D., Sheikh, Y. H. (eds) Towards new e-Infrastructure and e-Services for Developing Countries. *499*,391–413, Springer,Cham. https://doi.org/10.1007/ 978-3-031-34896-9_24

[10] [10] Diyaolu, I. J. 2021. Adoption of Sustainability in Clothing and Textile Production among Developing Countries. Journal of Environment and Sustainable Development, 1(1), 49-57. DOI:10.55921/WFZA5978

[11] Aggarwal, P."Fashion Product Images Dataset," Retrieved from https://www.kaggle.com/datasets/ paramaggarwal/fashion-product-images-dataset, Accessed: April 12, 2023.

[12] SJ, "Amazon Reviews on Women Dress (23k Datapoints)," Retrieved from https://www.kaggle.com/datasets/surajjha101/myntra-reviews-on-women-dresses-comprehensive, Accessed: April 12, 2023.

[13] Ololo, "Clothing Dataset (Full, High, Resolution)," Retrieved from https://www.kaggle.com/datasets/agrigorev/clothing-dataset-full, Accessed: April 12, 2022.

[14] Zalando Research, "Fashion MNIST," Retrieved from https://www.kaggle.com/datasets/zalando-research/fashionmnist, (April 18, 2023).

[15] Suthar, H. 2023. "Myntra Fashion Product Dataset," Retrieve from https://www.kaggle.com/datasets/hiteshsuthar101/myntra-fashion-roduct-dataset, (April 18).

[16] Ni, J., Li, J., and McAuley, J. 2019. "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 188-197.

[17] Nair, S. Patil, K., Waghela, H. and Pansambal, S. 2022 "Outfit Recommendation–Using Image Processing," *Journal of Algebraic Statistics*, vol. 13, no. 2, pp. 1699-1706.

[18] Stan, C. and Mocanu, I. 2019. "An Intelligent personalized fashion recommendation system," in 2019 22nd International Conference on Control Systems and Computer Science (CSCS), IEEE, 2019, pp. 210-215.

[19] Yeruva, S., Sathvika, A., Sruthi, D., Reddy, D. Y., and Gopi, G. 2022. "Apparel Recommendation System using content- Based Filtering," International Journal of Recent Technology and Engineering, vol. 11, no. 4, pp. 46-51.

[20] L. Leininger, J. Gipson, K. Patterson, and B. Blanchard, 2020. "Advancing performance of retail recommendation systems," SMU Data Science Review, vol. 3, no. 1, p. 6, 2020.

[21] Sridevi, M., ManikyaArun, N. M. Sheshikala, M. and Sudarshan, E. 2020. "Personalized fashion recommender system with image-based neural networks," IOP Conference Series: Materials Science and Engineering, vol. 981, no. 2, p. 022073, IOP Publishing.

[22] Orisadare, E. A. 2023. "Fashion Product Image & Text Dataset – DMFRM-202k," Retrieved from https://www.kaggle.com/datasets/ayooluwaemmanu-el/fashion-product-dataset-cmfrm-202k, (May 5)