Recommendation System Building based on CNN and TF-IDF Approaches

Mingshen Li *

College of Physical Science, University of Aberdeen, Aberdeen, The United Kingdom

* Corresponding author: t02ml22@abdn.ac.uk

Abstract. Online shopping has becoming increasingly popular. Customers remain engaged with shopping sites due to effective product recommendations. By analyzing users' past actions and preferences, shopping platforms effectively identify customer interests and desires, enhancing their overall shopping experience. This paper presents a recommendation system that displays 5 items relevant to the product currently being viewed by the customer. The system processes images with an adapted ResNet-50, a deep learning image classification model in convolutional neural network (CNN) and use the acquired embedded vector to determine similarities with cosine functions. Additionally, the system employs term frequency-inverse document frequency (TF-IDF) method for text analysis in product descriptions, generating word embedding that assists recommendations. This blend of visual and textual analysis ensures that the suggestions closely match the item's category. The system achieves 94.72% accuracy in subcategories' classification, a 4% increase compared to using the CNN method alone, confirming its effectiveness in recommending relevant items.

Keywords: Recommendation system, deep learning, image classification, text analysis.

1. Introduction

Over the past decade, growing internet usage has significantly boosted the online shopping market. The spread of the coronavirus disease has provided a further push to e-commerce activities as customers shift toward online platforms to follow social distancing protocols [1]. In developed economies, online retail spending has seen significant growth during the first year of the pandemic, with countries such as Britain at 19%, Australia at 7.2%, USA at 10%, France at 14%, and Germany at 17% [2]. The customers are attracted to online shopping for its recreational aspects, compelling deals, and more importantly, the opportunity to compare quality and prices for optimal bargains [3-4]. A proven successful sales strategy involves grouping similar products together, thereby providing customers with a selection of relevant alternatives influenced by frequently purchased items. Traditional online marketing falls short in this aspect. A lack of chance to interact with actual shelves makes it challenging for customers to compare more product at a time. Recommendation systems, by effectively addressing these issues, have become increasingly popular on websites such as eBay and the Korean e-commerce site SK Planet. Series of studies found that the relationship between "visual intents" such as color, texture, material, and design, and user feedback in the form of clicks, likes, and purchases, reveals that visual relevance not only positively correlates with but can also influences user feedback [5-7]. On the other hand, text queries continue to retain a significant user base, offering a unique perspective on user needs [8]. Combining both methods could lead to a more precise and effective recommendation system. A new approach emerges to enhances visual classification with textual data analysis, thus boosting the effectiveness of visual-only method in e-commerce environments.

The study proposes a recommendation system aimed at delivering suitable product suggestions by analyzing users' past behaviors and preferences, such as page visits and search history, to produce buying suggestions that closely match their interests. In this study, customer actions are defined as one item they currently viewing, represented by one of the images taken from Fashion Product image set [9]. The system utilizes a CNN architecture to identify similarities in product images and incorporates text analysis to improve the precision of recommendations. After constructing the initial

system, the study focuses on enhancing the model's precision. The results show that compared with traditional recommendation systems, the recommendation system constructed in this study performs well, can simultaneously improve customer satisfaction and merchant sales revenue, and can achieve a win-win situation between merchants and customers.

2. Experiment data

2.1. Data Description

The Fashion Product images dataset, created by Param Aggarwal in 2018, is a key resource for machine learning and image recognition in the apparel industry [10]. It features 44,000 high-resolution images (2400x1800 pixels) of varied fashion products against white or gray backgrounds, labeled with numerical IDs for categorization. The dataset is divided into two segments: the image collection and an accompanying csv file with 44,000 records, each corresponding to an image ID. This file encompasses 10 attributes in each product for detailed analysis, including gender, master categories, subcategories, article type, color, season, year, usage, and product name. These attributes facilitate a dual analysis approach: visual processing through CNN for image similarity and natural language processing (NLP) techniques for textual understanding, thereby determining similar items for the recommendation system. Data integrity is maintained by manual input from researchers, ensuring precision in linking images to their textual data, thus streamlining classifier training processes.

2.2. Data Preprocessing

The dataset originally comprises 44,000 high-resolution images and is valuable for its detailed product visuals. However, the deep learning model used a 50-layer neural network, which does not always necessitate such high-resolution. For enhanced computational efficiency, the study selects the initial 5,000 images from the set, re-sizing them to 240x180 pixels, thereby diminishing the data volume to just 0.11% of its original size. During the data cleaning process, rows causing errors are removed, and the dataset is modified to ensure each image is correctly linked with its corresponding ID in the CSV file. The index of the dataset is also reset for consistency. The next step in the study involves text data preprocessing, which starts by merging headlines and descriptions into one column. The combined text is then processed through several stages: removing non-alphanumeric characters, converting all text to lowercase, breaking down sentences into individual tokens, eliminating irrelevant stop-words, and rewriting words to their root forms. This comprehensive preprocessing prepares the text for detailed analysis.

3. Methodology and Procedures

3.1. Convolutional Neural Networks

In recent years, deep learning with a focus on Convolutional Neural Networks (CNNs) has emerged as a crucial technique in image classification. Developed by Yann LeCun in the late 1980s and early 1990s, CNNs mark a shift from traditional methods reliant on manual filters [11]. They excel in detecting complex patterns, with layers for local pattern recognition, data down-sampling while retaining key information, and classification based on extracted features.

The notebook employs the ResNet-50 CNN, featuring 50 layers and designed to tackle the vanishing gradient problem in deep networks using "skip" connections [12]. This architecture enables training deeper models without performance loss. ResNet-50 is enhanced with pre-trained weights from the ImageNet dataset, containing millions of images across 1,000 classes, broadening its feature recognition range. The recommendation algorithm starts with ResNet-50, adapted to transform 5,000 resized dataset images into one-dimensional vectors. This vector representation makes comparing image similarities more efficient, and enables more precise similarity assessments through advanced

mathematical techniques. Notably, the final fully connected layer of ResNet-50 is omitted, as its original purpose for classifying images into 1,000 categories is unsuitable for vector output tasks. Instead, pooling layers are incorporated to attain a uniform vector representation for every image, known as "embedding". This approach facilitates more accurate recommendations by focusing on image-based similarities.

3.2. Embedding

Embedding transforms raw data into a format more suitable for analysis, enhancing pattern recognition and decision-making. This vector-based representation greatly enhances the efficiency of mathematical operations, especially for sophisticated tasks like similarity calculations and clustering. In this study, images are first processed by the adapted ResNet-50 CNN architecture, followed by the application of an advanced pooling technique for embedding, ultimately leading to the formation of new vectors. This process transforms each image into a fix-sized, one-dimensional vector. The resulting embedded data frame features 2,048 columns, each a numerical feature of the image data, a marked increase from the original data frame's 10 features, such as color and season, thus significantly expanding the dimensionality.

3.3. Cosine Similarity

Cosine similarity is a metric that defines document similarity based on the angle between vectors in an inner product space using the cosine function. It measures the cosine of the angle between two vectors projected in a multi-dimensional space.

$$cos(\theta) = \frac{\Sigma_{i=1}^{n} A_i B_i}{\sqrt{\Sigma_{i=1}^{n} {A_i}^2}\sqrt{\Sigma_{i=1}^{n} {B_i}^2}} \tag{1}$$

The above equation represents the cosine similarity between two non-zero vectors, A and B. The equation is in a fracture form. The numerator is the sum of products of A and B's matching elements. The denominator consists of the square roots of the sum of squares of each vector's elements, representing the magnitudes of vectors A and B. Dividing the dot product by the product of the vectors' magnitudes provides a measure of their directional similarity.

Cosine similarity is widely used in recommendation systems where understanding the relative orientation of vectors is more important than comparing Euclidean distances between them. The study uses embedding to transform image data into one-dimensional vectors, utilizing cosine similarity to measure similarity between any 2 of these vectors. Upon receiving an image as input, the system calculates its similarity to other images and selects the top 5 vectors with the highest similarity scores. These 5 images are then displayed in a pre-designed grid layout. This setup enables the recommendation algorithm to effectively handle any input image.

3.4. Textual Method

TF-IDF, a statistical measure coined by K.S. Jones in the 1970s, evaluates word relevance by balancing its frequency against its prevalence in a corpus [13]. This approach blends term frequency (TF) and inverse document frequency (IDF) to proficiently assign weights to words during information retrieval and text mining.

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \tag{2}$$

Here, $tf_{i,j}$ represents the number of occurrences of the word i in document j, dfi is the number of documents containing word i, and N is the total number of documents in the collection. The TF component calculates the ratio of a word's occurrence in a document to the total number of words in that document, increasing as the word appears more frequently. The IDF component evaluates the

weight of rare words across all documents in the collection, with rarer words receiving higher scores. Combining TF and IDF, the TF-IDF score (denoted as w) for a word in a document is derived. This score, being the product of TF and IDF, reflects the importance of a word in a document, adjusted for its commonness or rarity across the collection.

In this study, TF-IDF was used to convert product descriptions into numerical vectors, capturing the significance of words within each product's context. These vectors were then compared using a similarity metric, creating a map that reflects how textually similar each product is to every other one. The study ranks these similarities to identify the top matches for a given product. Essentially, it uses TF-IDF for understanding text content and similarity measures for ranking, facilitating a recommendation process based on textual resemblance between products.

3.5. Combination

In this study, the integration of visual and textual similarity rankings is conducted by first identifying items common to both methods. When the intersection between the image-based and text-based methods yields more than n items, in which n is the total number of recommendations, the top nth recommendations are chosen, . This selection is based on the order of similarity scores determined by the image-based method, prioritizing the highest scoring items. If the intersection yields fewer than n items, the study supplements these with additional items from the textual method's top recommendations, aiming to complete a list with exactly n items. This approach ensures a balanced and comprehensive recommendation set, combining insights from both visual and textual analysis.

4. Results

4.1. An Example Tested by Human Vision

A successful recommendation system excels in providing relevant and personalized suggestions to users. The effectiveness of these suggestions will be evaluated visually. In the upcoming trail, one items are chosen at random as inputs. The algorithm then generates six recommended images (n=6) for each item. By drawing on their shopping experiences, individuals will determine if the recommended items genuinely align with their preferences when given the initial item. The displayed image showcases seven male models wearing grid-patterned shirts, with the top image labeled "im251" as the reference item. Below it is six recommended items, labeled from 251 to 3349. The key similarity among these shirts is their grid-like pattern, indicating that the recommendation system not only identifies the item by shape but also focuses on the pattern for related suggestions. This mirrors a typical shopping experience where buyers compare similar styles, demonstrating the effectiveness of the recommendation algorithm (See Fig. 1).

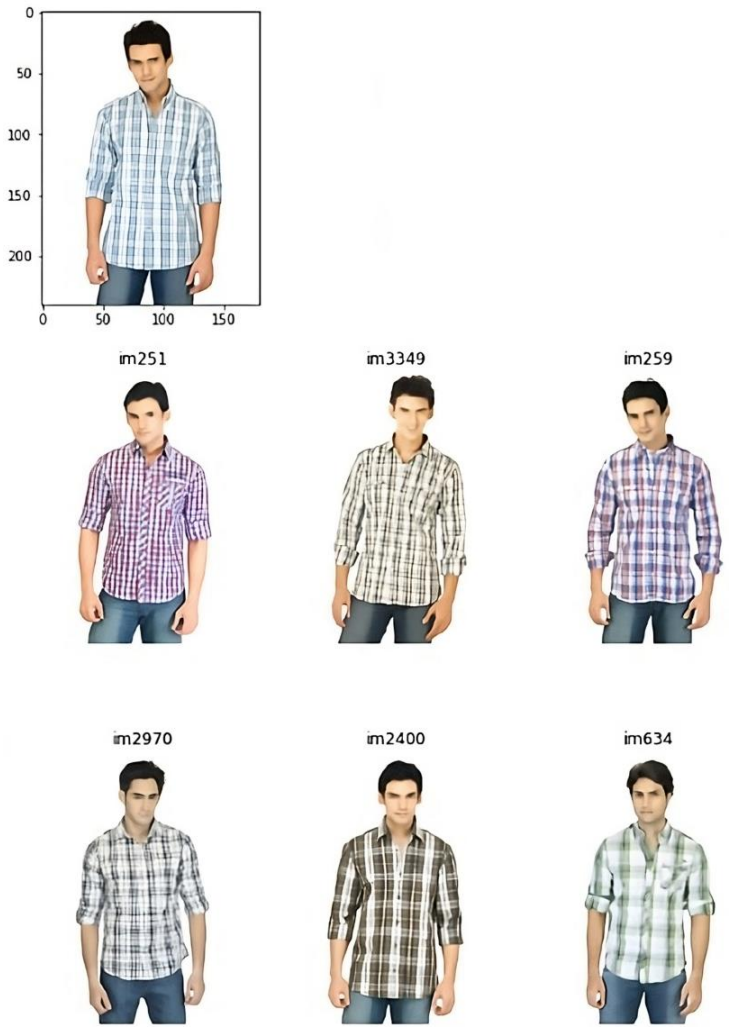


Figure 1. A recommendation system with 1 item as input and 6 items as recommendations

4.2. Finding Recommendation Number

The experiment's key objective is to identify the ideal number of recommendations, n. Setting n too high may lower accuracy as it includes less relevant items further down the similarity ranking, while too low an n risks providing an inadequate range of options. The system's accuracy is evaluated with a multi-class confusion matrix, which details correct and incorrect predictions using the masterCategory and subCategory data columns. This information is annotated manually in the dataset and are independent of the visual and text similarities previously examined. The experiment seeks the optimal recommendation count, n, starting with n=1 and incrementally increasing to n=9. Their accuracy at each value of n helps ascertain the most effective number of recommendations (See Fig. 2).

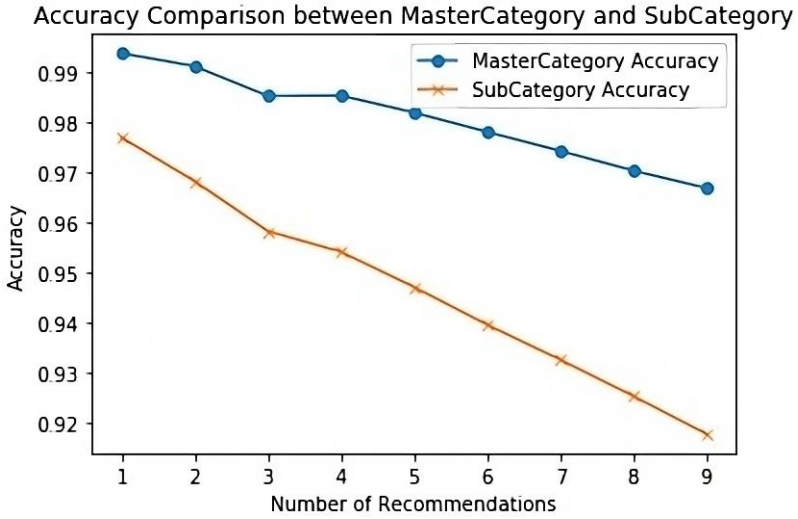


Figure 2. Accuracy of master categories and subcategories vs. the number of recommendations

The graph shows a decline of accuracy for both master categories and subcategories as the number of recommendations increases. Given that the subcategory encompasses a broader range of possible values, its accuracy is consistently lower than that of the master category at each recommendation level. To ensure the recommendation system's effectiveness, the study sets a manual threshold requiring at least 98% accuracy of the master categories, thus determining that an n value of 5 meets this criterion while still providing a sufficient number of recommendations.

4.3. Refine System and Show in Confusion Matrix

Adjust the algorithm to set the optimal recommendation number at 5 and refine the system accordingly. The provided example above demonstrates the system is functioning as intended. Its performance is examined by presenting a multi-class confusion matrix and calculate the accuracy. This visualization aids in detailing correct and incorrect predictions, offering insights into the type and frequency of errors (See Fig. 3, Fig. 4, and Fig. 5).

Figure 3. Refined recommendation system with 1 item as input and 5 items as recommendations

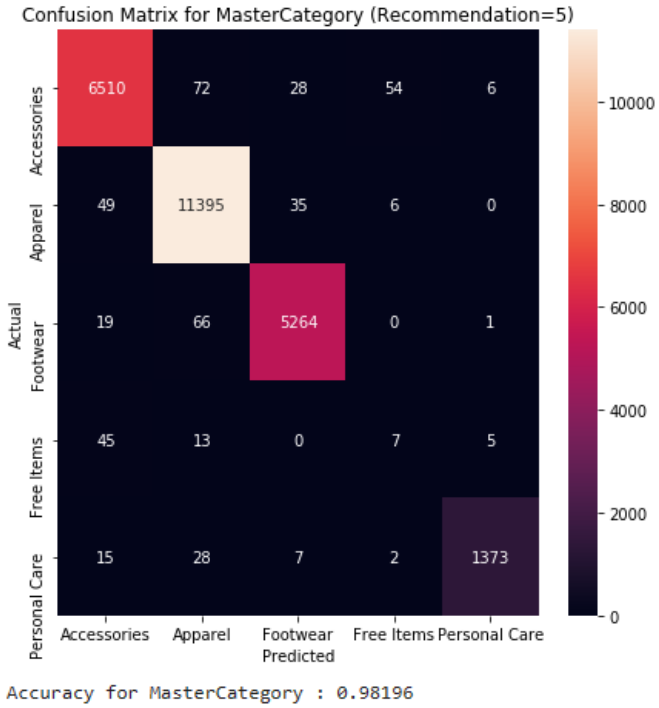


Figure 4. Confusion matrix of master categories for refined algorithm

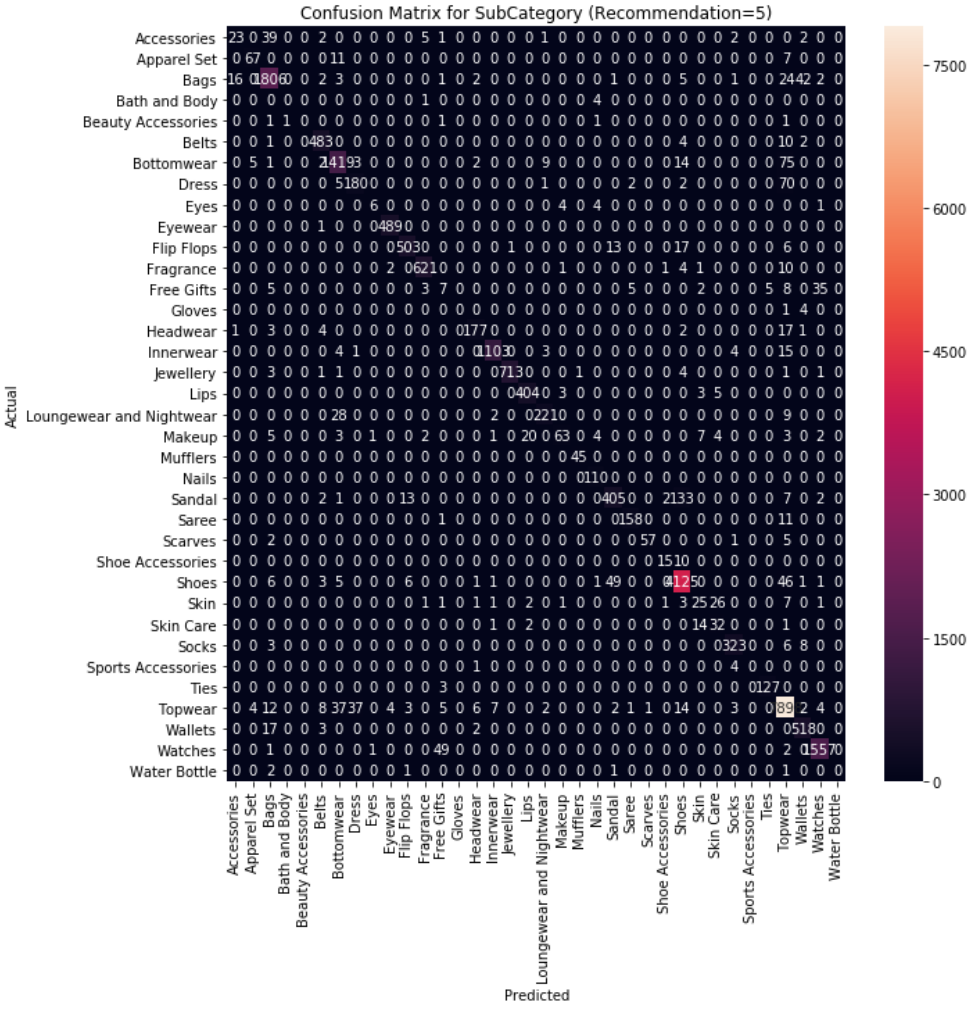
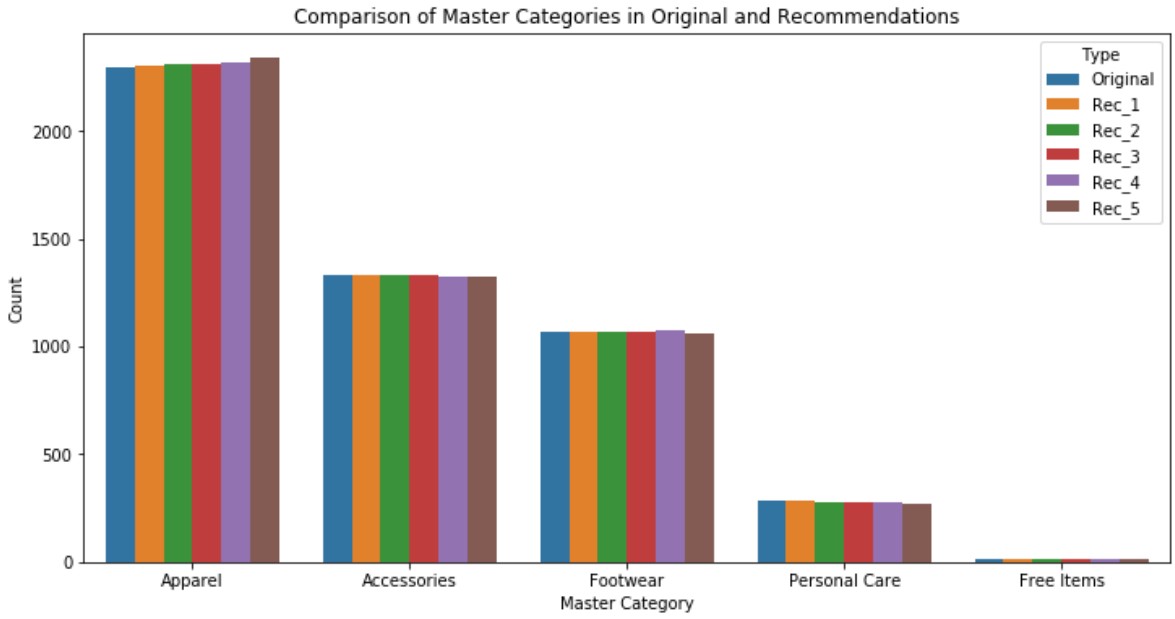


Figure 5. Confusion matrix of subcategories for refined algorithm

For recommendation number n=5, the accuracy on measuring master categories is 98.20%, and the accuracy on measuring subcategory is 94.72%.

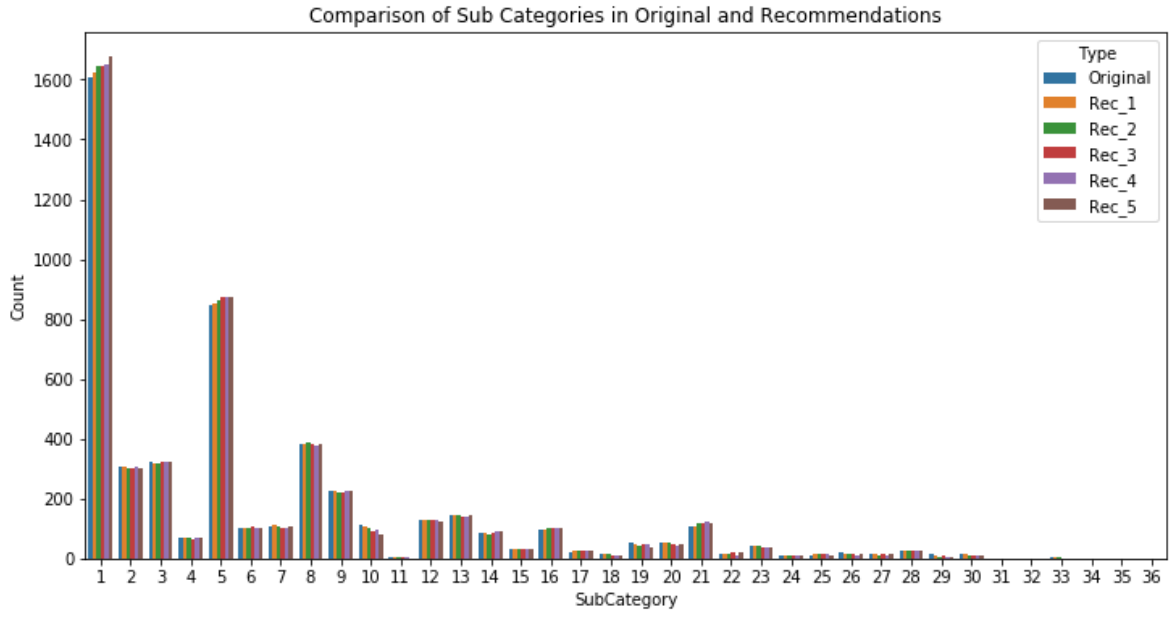
4.4. Evaluate the System in Bar Chart

The stacked bar chart illustrates the distribution of master categories in the original dataset compared to 5 recommended items based on the original items. The alignment of the master categories' distribution between the recommended and original items demonstrates the convincing performance of the recommendation system. Additionally, a comparison of subcategory distributions in the chart below further validates the effectiveness of the recommendation system (See Fig.6 and Fig. 7).




```
Accuracy for MasterCategory : 0.98196
```

Figure 6. Stacked bar chart of master categories with refined algorithm




```
Accuracy for SubCategory : 0.9472
```

Figure 7. Stacked bar chart of subcategories with refined algorithm

4.5. Compare the System with Visual-only Method

The study emphasizes the effectiveness of combining visual-based and textual-based methods over using the visual-based method alone. For comparative analysis, the study presents a stacked bar chart alongside the accuracy metrics obtained using only the visual method. This comparison aims to highlight the enhanced performance of the integrated approach (See Fig. 8 and Fig. 9).

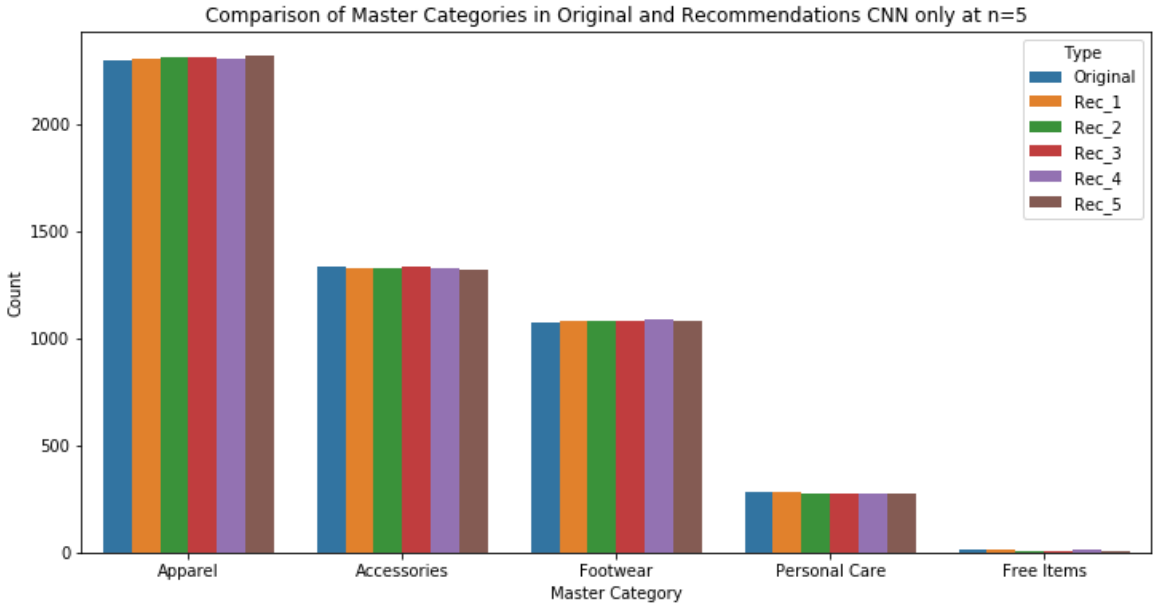


Figure 8. Stacked bar chart of master categories using visual-only method

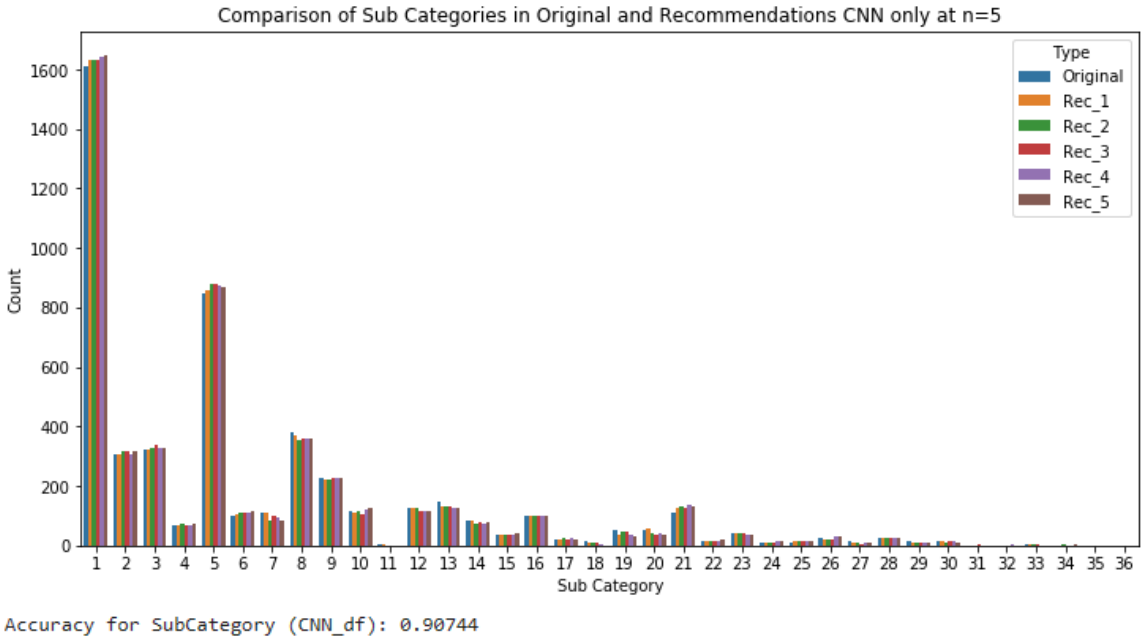


Figure 9. Stacked bar chart of subcategories using visual-only method

The results demonstrate a 98.49% accuracy of the master categories and a 90.74% accuracy of the subcategories. This indicates that the combined method ties the accuracy of the visual-only approach for master categories and significantly improves subcategories' accuracy by 4%.

5. Conclusion

The algorithm effectively suggests 5 additional items for consumers based on an input item. This number of recommendations is determined by a manually-set threshold, ensuring at least 98% accuracy in measuring with the general category. Its performance is validated by multiple methods including human vision, confusion matrix with accuracy, and stacked bar charts. Moreover, the combined method matches the visual-only approach in master categories' accuracy and notably enhances subcategories' accuracy. Implementing this system in online platforms is expected to be highly effective.

Future enhancements will consider the shopper's browsing behavior, expanding recommendations to include not just same-category items, but also a wider variety of options like diverse brands, complementary accessories, or alternative categories. This evident in cases where customers occasionally look for distinct products like batteries for an electronic toy. Future research may integrate user feedback, click-through rates, and time spent on recommended items provide valuable insights into user satisfaction and engagement. Such an enhanced method goes beyond recognizing similarities but could interpret customer's shopping interests, thus enhance their shopping experience.

References

[1] Zubenko, Y. GLOBAL E-COMMERCE DEVELOPMENT AND ITS IMPACT ON INTERNATIONAL MARKETS. Організаційний комітет, 2023, 305.

[2] Mofokeng, T. The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience. Cogent Business & Management, 2021, 8.

[3] Alavi, S.A., Rezaei, S., Valaei, N. et al. Examining shopping mall consumer decision-making styles, satisfaction and purchase intention. The International Review of Retail, Distribution and Consumer Research, 2016, 26: 272 - 303.

[4] Pandey, S., Chawla, D. Online customer experience (OCE) in clothing e-retail. Exploring OCE dimensions and their impact on satisfaction and loyalty – Does gender matter? International Journal of Retail & Distribution Management, 2018, 46 (3): 323 – 346.

[5] Togashi, R., Sakai, T. Visual Intents vs. Clicks, Likes, and Purchases in E-commerce. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020.

[6] Liao, L., He, X., Zhao, B. et al. Interpretable Multimodal Retrieval for Fashion Products. Proceedings of the 26th ACM international conference on Multimedia, 2018.

[7] Cho, B., Potluri, R.M., Youn, M. et al. A Study on the effect of product recommendation system on customer satisfaction: focused on the online shopping mall. Journal of Industrial Distribution & Business, 2020.

[8] Martínez, G., Saavedra, J.M., Murrugara-Llerena, N. VETE: improving visual embeddings through text descriptions for eCommerce search engines. Multimedia Tools and Applications, 2023: 1 - 37.

[9] Marlesson. Building a Recommendation System Using CNN - v2. Retrieved from https://www.kaggle.com/code/marlesson/building-a-recommendation-system-using-cnn-v2.

[10] Aggarwal, P. Fashion Product Images Dataset. Retrieved from https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset

[11] LeCun, Y., Boser, B.E., Denker, J.S. et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1989, 1: 541 - 551.

[12] He, K., Zhang, X., Ren, S.et al. Deep Residual Learning for Image Recognition, 2015, ArXiv, 1512. 03385.

[13] Abubakar, H.D., Umar, M. Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 2022.