# Apparel Recommendation System

Shivendra Pratap Singh[1], Vishnu Mishra[2] , Sunil Yadav[3], Apoorv Mishra[4]

Department of Computer Science and Engineering, 1 Maharana Pratap Engineering College, Kanpur, Uttar Pradesh, India

Maharana Pratap Engineering College Kanpur, Uttar Pradesh, India

---

**Abstract:** Nowadays, people are constantly moving towards various fashion products as a result the e-commerce market for garments is growing rapidly. Online stores must update their features according to user requirements and preferences. However, there are too many options for users to select from these online stores which may leave them in a dilemma to identify the correct outfit, save the user time, and increase sales, efficient recommendation systems are becoming a necessity for online retailers. In this paper, we proposed an Apparel Recommendation System that generates recommendations for users based on their input. We used a real-world data set taken from the online market giant Amazon using Amazon's Product Advertising API. We aim to use keywords like brand, color, size, etc., to recommend. Data exploration to get detailed information about our dataset, Data Cleaning(pre-processing) to remove invalid sections, Model selection (We have compared different feature extraction techniques like bag of words, TF-IDF, and word2vec model) to find out efficient techniques and Deployment of the model that could facilitate recommendation system to simplify the task of apparel recommendation system. The accuracy of the model is identified using the response time and content matching.

Keywords: Apparel, Recommendation, TF-IDF, Bag-of-Words, Content-Based-Filtering

# I. INTRODUCTION

In the present world, a user usually adopts a system that provides services to make the user's job easier. A recommendation system is one such feature that makes the user's experience comfortable and easy. A user exploring all the products on the website for shopping for a specific product is not an easy and possible task. Here Recommendation Systems play a very important role which helps the user find what they are looking for. If a user is searching for a shirt according to his preferences and finds a particular shirt that matches his choice, then if the shopping website recommends shirts that are alike to the chosen shirt it enhances the user's shopping experience. This will lead to happy customers which will also increase sales and customer engagement with the company. In this way, the recommender system benefits users in finding items of their interest. Recommender systems guide users to the items which users are most likely to purchase. Recommender systems are like a well-known service person who knows the user's history and makes his decision-making process easier. The user may feel known and start exploring more and more products which ultimately leads to more purchases. Companies focus on personalized product recommendations to increase their sales. Personalized product recommendations help companies to engage more with users and provide a delightful shopping experience. Personalized product recommendations help in finding items that are most relevant to the users. When a user can relate to the recommendations perfectly with his choices then he is bound to visit the website again and again and chances are high that he may also become a loyal customer. A recommendation system is an added benefit to the companies which they can utilize to move ahead of their competitors in this competitive world. Recommendation systems are of use in e-commerce websites and online streaming companies. These recommendation systems allow customers of a company to see relevant products. The ease of use of the products and being given more customer-related choices enhance the user experience. The recommender systems help in enhancing customer engagement and brand loyalty. They reduce the transaction costs of finding and selecting items in an online shopping environment. Recommender systems create a delightful user experience while driving incremental revenue for websites/companies.

Our work has a user web interface from which the user can provide a string containing the details of the product like color, product type, product brand, etc. as these are one of the main parameters based on which a user intends to get similar results of the product and also as the input user can provide several similar products, he wants to retrieve so that he may choose a wide range of products according to his available time and resources. The dataset is processed by the importance of features. ASIN number

in the dataset plays a key role in uniquely identifying a product. The text-based recommendation of the products makes the title a suitable feature to be selected for the data analysis. So, TF-IDF based prediction is the main model used for text-based prediction.

The output of the project is the listing of products whose listing is according to the given number of similar products and also their brief details. Recommendation systems are useful for customers and companies. They help in reducing transaction costs of finding and choosing items in an online shopping environment. Recommendation systems play a key role in improving revenues, and they are the major reason for selling more products [1]. The recommender system helps in increasing the average order values. This is because the website displays more personalized choices since the customers like to have the products that they strongly desire. Recommender systems also help in the increasing number of items per order. When the applications display the products which meet the user's interest, then the user is more likely to add those items to the bag [2]. The most successful company amazon reported that there is a 29% increase in sales because of recommender systems [3]. And it also specified that 35% of its revenue is generated by recommendation systems[4]. Netflix which has 182.8 million subscribers specified that 80% of the whole streaming time is achieved only by their recommender systems [5]. Flipkart stated that due to the recommender system there was a 10% increment in the CTR [6].

# I. DATASET

The dataset that we are using in the project is very exciting. It is extracted by web scraping the Amazon apparel product data through Amazon product advertising API. From most of the brands in the dataset, we are inferring that the dataset is taken from amazon.com which is a USA website of amazon ( www.amazon.com ).

## A. Features of dataset

The dataset consists of women's clothing products. It has 1,83,138 data points and 19 features. In this project, we limited ourselves to only 7 features based on certain criteria. The 7 attributes of the dataset that we are using in our project are:

**1.ASIN**- Amazon Standard Identification Number. It is a unique identifier of 10 letters/ numbers for a product that's assigned by Amazon. It is primarily used for product identification and also solves the problem of duplicate products by merging them. Amazon follows the rule that while a product exists in the amazon catalog the new seller needs to use an existing ASIN number which is the same for all the sellers who sell the same product and this is usually the case for resellers, products with wide distributions. A product that is not in the Amazon catalog will be assigned a unique ASIN number [15]. In this way the ASIN number is unique for every unique product, so we considered ASIN in our data analysis as it is unique for every product and it allows easy identification of product.

**2.Brand**- The brands that are on sale for Amazon are unique and there will not be any infringement or inaccurate brand information .

Amazon does not accept listings that do not have proper permissions for selling that product which comes under brand infringement and it also does not accept products in the branch field has generic naming like (Shirt) and unapproved brand [17]. So, the brands of apparel in amazon listings are accurate and they help in recommending similar brands' apparel which will help users to choose the products.

**3. Colour**- While listing a product on the amazon website color of the product also plays a key role because two products that are the same but differ in color of the product can be under the "Variation Parent" property [18]. A customer may be interested in similar color products or different color products from the same brand. The Colour of the product is a kind of more complex data because we can have a color composition that is too diverse, especially for apparel and color is too important for apparel listing because the amazon website enables its users the featured shop by colour [19]. So, the color of a product also helps in the recommendation.

**4. Medium image URL-** Amazon imposes certain criteria on the images that have to be followed by sellers while listing a product. Images almost show the story as much as the reading of the description or details of the product. The medium-sized URL feature of data helps in retrieving the image and using it for interactive display and the user easily assesses the products.

**5. Product type name-** Amazon's classification system uses values supplied by sellers to determine where a product appears in the catalog and the product type of that particular product [20]. The product type name helps in grouping related. The user can recommend products according to his/her relevant products from the group of products.

**6. Title**-The title of a product is almost like a short description of the product. It conveys a lot of information about a product. The title of a product is considered as clean data about a product as Amazon imposes strict rules like the title should be of fixed length and there should not be any promotional offers or subjective commentary such as “hot item”, “fast selling” etc.. and should not contain non-language ASCII characters [21]. So, with all of these rules in the title section, it is easy to analyze and generate insights.

**7.Formatted price**- The price details of the product too help in recommending similar items. The prices of products that are featured on the amazon website are more dynamic because they use this feature of dynamic pricing as the main idea to adjust their prices at a rapid pace to meet the market demand than any other e-commerce website [22]. The too-complex pricing model makes over 250 million price updates. An average product price will change once every 10 minutes [23]. The formatted price has too many null rows/data points because Walmart has scraped the amazon data and got the amazon pricing for products and relevantly adjusted the product prices on the Walmart website. So, to remain in the competitive business amazon pricing data may not be revealed in web scraping.

# B. UNUSED FEATURES

The other attributes of the dataset which are not made into the model for a recommendation of products are:

**8. SKU**- SKU which stands for stock keeping unit is a number that can be created by a seller (according to to sell.amazon.in). The seller may mark a batch of products with the same SKU number which may lead to confusion also in the dataset we can observe that most of the SKU numbers are not provided. On the other hand, the ASIN number stands as a great alternative for all situations where SKU cannot be relied on.

**9. Availability and availability type**- Availability and availability of a product do not give any necessary details for the apparel recommendation project because we are addressing only the recommendation part of apparel and not the selling or business aspects of the online apparel system.

**Reviews, editorial review and editorial review**- Reviews do not give much information to process in the apparel recommendation system as the approach for the recommendation system remains content-based filtering, not collaborative filtering. So, we are not going to consider the approach of a user towards a product in the recommender systems. And we are not going to use the sentiment analysis part in this model.

**Author-** The author feature in the dataset has no values at all.

**Publisher and Manufacturer-** While a user searching for a product he may consider products from the same brand but surely not just by considering only the manufacturer. A manufacturer of apparel can be a third party who manufactures a product on behalf of a company but not manufactures only a specific brand of apparel and the publisher stands as the third-party person who may sell different brands of apparel who sells products on behalf of the company

**Large image URL and small image URL-** As we have already considered medium image URLs for the visual representation of the recommended results, we did not consider large image URLs and small image URLs.

**Model-**The model represents the person who is featured in photos with the model wearing apparel and the model information does not convey much useful information for the recommender systems so it is not useful in the recommendation of apparel.

# C.DATA PRE-PROCESSING

The dataset that we have taken is of size ~183k records. And by eliminating unwanted data items and by performing text pre-processing, and deduping we reduced it to ~16k. Initially, we eliminated the data items that have a price as a null value because the price is the main option that is considered by any average consumer while selecting or searching for a product. On eliminating those data items which do not have price values we are left with 28,395 records as shown in figure-1. Then we eliminated the data items that are having color values as null in other words which are not having any price details.

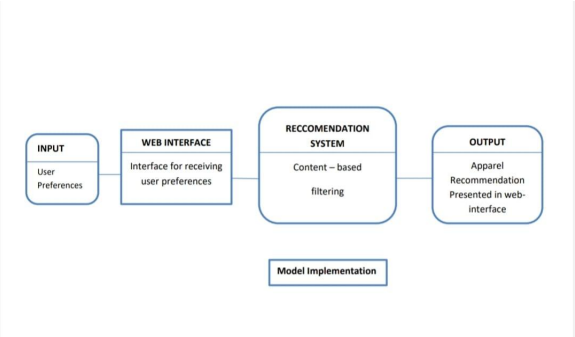
The dataset that we have taken is of size ~183k records. And by eliminating unwanted data items and by performing text pre-processing, and deduping we reduced it to ~16k. Initially, we eliminated the data items that have a price as a null value because the price is the main option that is considered by any average consumer while selecting or searching for a product. On eliminating those data items which do not have price values we are left with 28,395 records as shown in figure-1. Then we eliminated the data items that are having color values as null in other words which are not having any price details.

On eliminating those data items which do not have color values we are left with 28,385 records. The below snippet of code describes the process that is done for eliminating null price and null color items.

1. **Removal of null-valued rows**
2. **Removal of short title rows**
3. **Duplicate title rows removal**
4. **Removal of titles that differ in less than 2 words**

# METHODOLOGY

The entire work of this paper is represented with the following architectural model diagram as shown in figure-5. This model takes user preferences as input in the web interface. This input will further be used to process and produce the best possible apparel recommendations. The user can give a string as the input to the model which may denote color, product type, or brand of product, as these can be the parameters on which a user intends to get similar recommendations and also can give several recommendations the user wants to retrieve.




The text-based recommendation of the products makes the title a suitable feature to be selected for the data analysis. So, the bag of words model, TF-IDF based prediction, IDF-based prediction, and Word2vec models are the main models used for text-based prediction. The output of the project is the listing of products whose listing is according to the given number of similar products and also their brief details. Finally, the Output of the project is apparel recommendations which will be presented in a web interface recommended by the model considering the preferences given by the user.

## A. Bag of Words

The bag of Words model is used to convert text into some numerical format. It stores the frequency of the most frequent words. In the Bag of Words implementation of our model, we started with the preprocessed data where the titles don't have any stop words. We extracted a bag of words by taking all the product titles. For each title, we get a d-dimensional vector in which we have a count of occurrence of each word in that particular title. This is obtained with the help of the CountVectorizer function which is imported from the sklearn library. In this way, we get vectors for all 16042 titles. By running the fit transform function we get a matrix that has titles as rows and each of its words as columns. In the bag of words function, we are going to calculate the pairwise distances which is nothing but the Euclidean distance of title T with all of our points. And then we sort those distances accordingly, from that sorted array we are going to retrieve the required no of titles which has the smallest pairwise distance.

## B. TF-IDF

TF-IDF model is an analytical measure that evaluates how relevant a word is to a document in a collection of documents [24]. TF-IDF is a combination of two extraction methods. They are TF - Term Frequency and IDF - Inverse Document Frequency.

$$TF\text{-}IDF(x) = TF(x) * IDF(x) \quad (1)$$

TF-Term Frequency

IDF- Inverse Document Frequency

Term frequency calculates how frequently a word occurs in a document. TF(x)=count of the x in the document / total no of words in the document

Inverse Document Frequency calculates how common or rare a word is in the entire document set by which the importance of a word is known.

$$IDF(x) = \log(N / df + 1) \quad (2)$$

N-Total number of documents

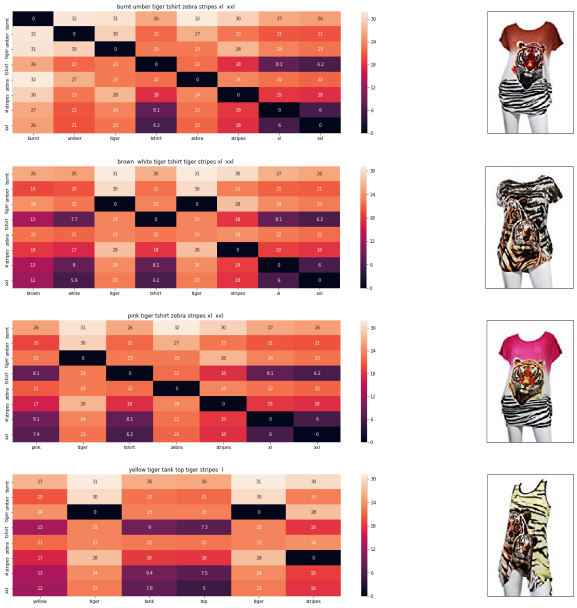
df-number of documents with term x.

In our project, we used the pre-built function Tfidf Vectorizer () offered by python's library sklearn. feature_extraction.

our raw data into a matrix of TF-IDF features. Min_df is used to remove the terms that appear too infrequently. By default, it is set to 1 which ignores the terms that appear in less than one document. By setting it to 0 we didn't remove any data because the data was already cleaned and pre-processed. Next, we used the same fit transform function used in Bag of Words to get a sparse matrix that has titles as rows and each of its words as columns. Now that we had required a sparse matrix we implemented our tfidf_model () function. In the tfidf_model(), we used pairwise distances() from python's sklearn.

## RESULT

The input given by the user is an animal pattern shirt. As mentioned before, the model is trained to retrieve the best-fit recommendations for the given input. In this instance, the model must return shirts of different kinds of animal patterns or a shirt that consists of animal pictures on it. We can observe that the major part of the shirts retrieved is zebra, tiger, and lion patterned or shirts with pictures of animals as shown in Fig. 6, 7, and 8



## CONCLUSION

In this paper, we were able to learn the usage of Bag of words and TF-IDF algorithms on the Amazon Apparel data. Firstly, we implemented the model using a bag of words algorithm. Based on the results, it was clear that the recommendations obtained are not desirable for the given input as words like shirt and zebra-pattern are given equal importance. Hence, we moved on to the TF-IDF algorithm which overcame the drawbacks of a bag of words. TF-IDF gives weightage to the words according to their uniqueness and importance of the word in each sentence which eventually resulted in desirable recommendations. We learned how to overcome the drawbacks of Bag of Words and make the recommendations more relevant.

## REFERENCES


**1.** Isinkaye, Folasade Olubusola, Y. O. Folajimi, and Bolande Adefowoke Ojokoh. "Recommendation systems: Principles, methods, and evaluation." *Egyptian Informatics Journal* 16.3 (2015): 261-273. [CrossRef]

**2.** Röllecke, Felix Johannes, Arnd Huchzermeier, and David Schröder. "Returning customers: The hidden strategic opportunity of returns management." California Management Review 60.2 (2018): 176-203. [CrossRef]

**3.** Amazon's Recommendation Engine: The Secret To Selling More Online Available Online http://rejoiner.com/resources/amazon-recommendations-secret-sellingonline (accessed on June 2021)

**4.** How Amazon Has Reorganised Around Artificial Intelligence and Machine learning Available Online: https://www.forbes.com/sites/blakemorgan/2018/07/16/how-amazon-has-re-organized-around-artificial-intelligence-and-machine-learning/?sh=250a76e77361 (accessed on June 2021)

**5.** Chong, D. "Deep dive into Netflix's recommender system." (2020).

**6**. Eisner, Alan, et al. "FLIPKART: WINNING IN INDIA?" *GLOBAL JOURNAL OF BUSINESS PEDAGOGY* 4.1 (2020): 79.**7.** Zhang, Qian, Jie Lu, and Yaochu Jin. "Artificial intelligence inrecommender systems." Complex & Intelligent Systems 7.1 (2021): 439-457. [CrossRef]

**8**.Prando, Allan Vidotti, and Solange N. Alves de Souza. "Modular Architecture for Recommender Systems Applied in a Brazilian e-Commerce." *JSW* 11.9 (2016): 912-923. [CrossRef]

**9**. Hwangbo, H., Kim, Y.S. and Cha, K.J., 2018. Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, *28*, pp.94-101. [CrossRef]

**10.** Hsiao, P.-C., Cheng, Y.-H., & Chen, K.-J. (2020). A personalized recommendation system for fashion e-commerce based on customer feedback. Electronic Commerce Research and Applications, 40, 1

**13.** Pandit A, Goel K, Jain M, Katre N. A Review of Clothes Matching and Recommendation Systems based on User Attributes.

**14.** Liu, Yu, et al. "Clothing recommendation system based on advanced user-based collaborative filtering algorithm." International Conference on Signal and Information Processing, Networking and Computers. Springer, Singapore, 2017. [CrossRef]

**15**. Read More What is Amazon ASIN number & how to get it? Available online: https://www.datafeedwatch.com/blog/amazon-asin-number-what-is-it-and-how-do-you-get-it (accessed on June 2021)

**16.** Amazon Brand Registry: Help Protect Your Brand on Amazon Available online: https://brandservices.amazon.com (accessed on June 2021)

**17.** Amazon Brand Name Policy Available online: https://sellercentral.amazon.in/gp/help/external/G2N3GKE5SGSHWY RZ (accessed on June 2021)

**18.** Creating color variation on the existing product - Listing Management & Reports - Amazon Seller Forums Available online: https://sellercentral.amazon.com/forums/t/creating-color-variation-on-existing-product/150470 (accessed on June 2021)

**19**. Navigation by size or color Available online: https://sellercentral.amazon.com/gp/help/external/G53001 (accessed on June 2021)

**20.** Classify your products with Product Classifier Available online: https://sellercentral.amazon.in/gp/help/external/G200956770 (accessed on June 2021)

**21.** Product title requirements Available online: https://sellercentral.amazon.in/gp/help/external/YTR6SYGFA5E3EQC (accessed on June 2021)

**22.** The Amazon Effect: Dynamic Pricing Done Right Available online: https://www.pragmaticinstitute.com/resources/articles/product/the-amazon-effect-dynamic-pricing-done-right/ (accessed on June 2021)

**23.** An Introduction to Amazon Pricing in 2021 Available online: https://www.omniaretail.com/blog/an-introduction-to-amazon-pricing (accessed on June 2021)

**24**. TF IDF from Monkeylearn https://monkeylearn.com/blog/what-is-tf-idf/ (accessed on June 2021)