# Introduction

Myntra is a larger Indian fashion e-commerce online shopping store and its headquartered in Bangalore, Karnataka, India. Ashutosh Lawani, Vineet Sexana, and Mukesh Bansal founded it in 2007. In 2007, Myntra started out as a business-to-business player. The company's primary focus in its early years was selling other businesses personalized gifts that could be customized on demand. It developed into a popular hub for customers looking to purchase personalized goods like t-shirts and mugs. 2011 saw Myntra abandon its personalization business and transition to fashion and lifestyle products. Flip Kart acquired Myntra in 2014. However, Myntra continuous to function as an independent company and did not merge with Flip kart.

One of the top online retailers is Myntra, which has a wide selection of clothing and accessories. It gives consumers access to a large selection of both domestic and foreign brands. Myntra offers a wide range of products, including sports and active wear, bags and bag packs, watches, sunglasses, accessories, footwear, apparel, and much more. Even though there are other online retailers selling clothing and lifestyle products, many customers Favor Myntra. at Myntra. One reason is the company offers the latest in fashion and some of the most unique products. Myntra also launched special programs such as ‘easy 30 days return and exchange’ and ‘try and buy’.

The Myntra Fashion Products is a large e-commerce product that contains information about over 1 million fashion products from the popular Indian e-commerce platform Myntra. The products include a wide variety of product attributes, such as product name, description, price, brand, category, and images. The products is a valuable resource for researchers and developers who are interested in fashion e-commerce, product recommendation systems, and image recognition.

Myntra is unique in that it's a fashion and beauty platform that focuses on making fashion accessible to all. Myntra offers a wide selection of products for different styles, budgets, and preferences. They also offer personalized recommendations, easy returns, and efficient customer service. Myntra is a one stop shop for all your fashion and lifestyle needs. Being India's largest e-commerce store for fashion and lifestyle products, Myntra aims at providing a hassle free and enjoyable shopping experience to shoppers across the country with the widest range of brands and products on its portal. The brand is making a conscious effort to bring the power of fashion to shoppers with an array of the latest and trendiest products available in the country.

The Myntra Fashion Products is a comprehensive collection of product information from Myntra, one of India's leading e-commerce platforms. This dataset comprises a wide range of products across various categories, including clothing, footwear, accessories, and home décor. It is a valuable resource for researchers and data scientists working on various aspects of e-commerce, fashion, and product recommendation systems.

This project report will explore the Myntra Fashion Products and use it to develop a product recommendation system. The system will be able to recommend products to users based on their past purchases and browsing history. The system will also be able to learn from user feedback and improve its recommendations over time.

# Literature Reviews

Number of studies has been done based on the Myntra Fashion Products has been used in a variety of research studies. Some of the associated work contains:

1. **"A Content-Based Recommendation System for Fashion E-commerce" by Hitesh Suthar (2021),** proposes a content-based recommendation system that uses the Myntra Fashion Product Dataset to recommend similar products to users with personalized product recommendations. The system extracts feature from product descriptions, images, and other attributes to create a comprehensive representation of each product. These representations are then used to compute the similarity between products and recommend products that are similar to those that the user has previously interacted with. The system was evaluated using the Myntra Fashion Product Dataset and achieved promising results, suggesting its potential for practical applications in fashion e-commerce.

2. **"Product Review Image Ranking for Fashion E-commerce" by Rahul Gupta et al. (2022),** proposes a method for ranking product review images on fashion e-commerce websites. The method leverages the Myntra Fashion Product Dataset to develop a ranking algorithm that considers factors such as image quality, relevance to the product, and user engagement. The researchers evaluated their method using the Myntra Fashion Product Dataset and demonstrated its effectiveness in improving the quality and relevance of displayed product review images. the method uses a machine learning algorithm to learn the importance of each of these factors. The algorithm is trained on the Myntra Fashion Product Dataset. Once the algorithm is trained, it can be used to rank product review images on other fashion e-commerce websites.

3. **"Fashion Clothing Products Dataset: Analysis and Insights" by Kiitan Olabiyi (2022)**, analyzes the Myntra Fashion Product Dataset to gain insights into fashion trends. The study explores various aspects of the dataset, including product categories, colors, materials, and pricing, to identify patterns and trends that can inform fashion designers, retailers, and consumers alike. The analysis reveals insights into popular fashion choices, seasonal trends, and emerging styles, providing a comprehensive understanding of the current fashion landscape. The findings of this study can be utilized to guide product development, marketing strategies, and consumer purchasing decisions, ensuring that fashion offerings align with current trends and consumer preferences.

## Objective

1) To identify the trends and patterns in Myntra Fashion Products.
2) To develop a natural language processing (NLP) model to extract keywords from product description.
3) To develop a natural language processing (NLP) model to clustering the product description and Individual category into coherent and similar groups.
4) To build a recommendation system that recommends complementary fashion product to customers.
5) To develop a multi-class sentiment classification model that classifies product ratings as positive, negative, or neutral.

# Methodology

As we decided to work on Myntra Fashion Industry. We decided to Study Different Brands of Myntra fashion products across India, all information used in this project is collected from online source and various websites. we decided to work on different brands of Myntra fashion products that will give us option to choose the perfect, comfortable and luxurious cloths.

The methodology consists of several steps:

**1. Dataset Collection:**

The initial step involved the collection of the Myntra fashion product dataset, which serves as the foundation for this project. The dataset was obtained from Kaggle website. It contains information about various fashion products available on the Myntra platform, including attributes such as product category, brand, price, and customer reviews & ratings and product description.

**2. Preprocessing and data exploration:**

An exploratory data analysis (EDA) was conducted to gain meaningful insights from the datasets. This included:

1) Recognizing the Dataset: Examining the structure, size, and data types of the features.
2) Handling Missing Data: Recognizing and filling in the dataset's missing values.
3) Creating summary statistics through exploratory data analysis (EDA)

**3. Feature Engineering:**

Feature engineering was performed to enhance the dataset relevance and suitability for analysis. This involved creating new features, transforming existing ones, and addressing any outliers or missing values present during the EDA process and by using EDA process we can identify the fashion trends and patterns in the dataset and find out the popular brands among the peoples.

**4. Model Evaluation:**

The performance of the developed models was evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. Also, we can develop a model by using NLP technique to extract the keywords and cluster the similar product description.

**5. Clustering Analysis:**

To identify patterns and groupings within the dataset, a clustering analysis was performed. Unsupervised machine learning algorithms, such as K-Means clustering, were applied to segment products based on features like price, brand, category and individual category.

**6. Recommendation System:**

Systems of recommendations can be a powerful tool for companies. By providing users with relevant recommendations, businesses can increase engagement, sales, and customer satisfaction. To build a recommendation system we can use product ratings, price, brand names and others factors.

**7. Results and Interpretation:**

The findings from the analysis, including clustering results, sentiment analysis insights, and model predictions, were interpreted and presented comprehensively. Visualizations, tables, and graphs were used to communicate the results effectively.

**Here some overview of python libraries used for data analysis:**

• NumPy for linear algebra o Pandas for tabular data manipulation & processing, CSV file I/O

• Matplotlib for data visualization (scatterplots, bar charts, histograms, etc.)

• Seaborn for data visualization and statistical plotting (built on top of Matplotlib)

• Plotly for interactive data visualization

• Scikit-learn for data modelling and machine learning

## 4.1. Data Description

The Myntra Fashion Product dataset is a large, diverse, and high-quality dataset of fashion products from the Indian e-commerce platform Myntra. The dataset contains 184913 rows and 12 columns. The dataset contains a wide variety of product information, including product descriptions, categories, attributes, and prices. This dataset is a valuable resource for research in areas such as fashion recommendation systems, Cluster Analysis and fashion trend analysis.

Here is small overview of Dataset:

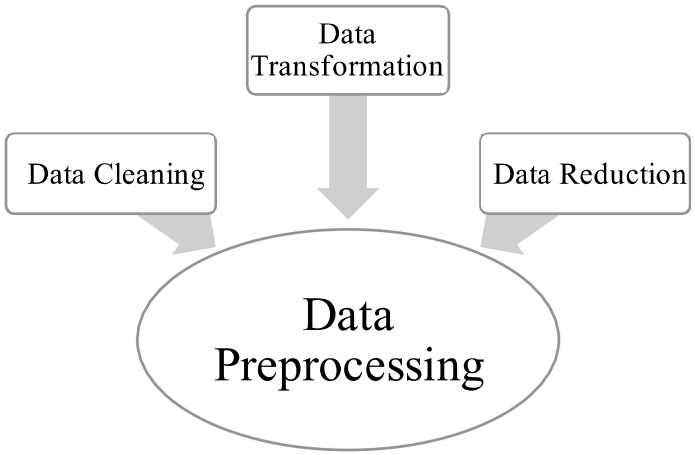
- Source: Kaggle.com
- Link: https://www.kaggle.com/datasets/manishmathias/myntra-fashion-dataset
- Myntra Fashion Clothing: December-2021 to January -2023

| | Product_id | BrandName | Category | Individual_category | category_by_Gender | OriginalPrice (in Rs) | Discount (in Rs) | Discount (in %) | Ratings | Reviews | SizeOption | Description |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2296012 | Roadster | Bottom Wear | jeans | Men | 1499.0 | 824.45 | 45 | 3.9 | 999 | 28, 30, 32, 34, 36 | roadster men navy blue slim fit mid rise clean... |
| 1 | 13780156 | LOCOMOTIVE | Bottom Wear | track-pants | Men | 1149.0 | 517.05 | 55 | 4.0 | 999 | S, M, L, XL | locomotive men black white solid slim fit tra... |
| 2 | 11895958 | Roadster | Topwear | shirts | Men | 1399.0 | 629.55 | 55 | 4.3 | 999 | 38, 40, 42, 44, 46, 48 | roadster men navy white black geometric print... |
| 3 | 4335679 | Zivame | Lingerie & Sleep Wear | shapewear | Women | 1295.0 | 893.55 | 31 | 4.2 | 999 | S, M, L, XL, XXL | zivame women black saree shapewear zi3023core0... |
| 4 | 11690882 | Roadster | Western | tshirts | Women | 599.0 | 389.35 | 35 | 4.2 | 999 | XS, S, M, L, XL | roadster women white solid v neck pure cotton ... |

## 4.2. Data Preprocessing

To clean and transform raw data for data mining, data preprocessing is a crucial step. It involves tasks like handling missing values, removing outliers, scaling features, and encoding categorical variables. By preprocessing the data, we ensure that data is in right format and ready for training our models. Data preprocessing is a primary step in machine learning where we prepare our data before training our models. It involves various techniques to clean, transform, and enhance the quality of our data.

Typical data preprocessing steps include the following:




- One important task in data preprocessing is handling missing values. We need to decide how to handle missing data, such as replacing them with the mean or median of the feature, or using more advanced techniques like regression or imputation.
- Encoding categorical variables is another important preprocessing step. Machine learning models typically work with numerical data, so we need to convert categorical variables into numerical representations. This can be done by using techniques like label encoding.
- Data preprocessing helps us in achieving better model performance and more accurate predictions. It ensures that our data is in the right format, free from inconsistencies, and ready to be fed into our machine learning models.

## 4.3. Data Cleaning

When it comes to data cleaning processes, there are several steps involved in ensuring that the data is accurate, consistent, and reliable. Here are some common data cleaning processes:

**Handling missing values**: This process involves identifying and dealing with missing data points in the dataset. Depending on the nature of the missing values, you can choose to either remove the records with missing values or impute the missing values using techniques like mean, median, or regression.

First step is checking missing values present in the dataset:

```
Product_id                  0
BrandName                   0
Category                    0
Individual_category         0
category_by_Gender          0
Description                 0
DiscountPrice (in Rs)   68053
OriginalPrice (in Rs)       0
DiscountOffer           18458
SizeOption                  0
Ratings                     0
Reviews                     0
Description.1               0
dtype: int64
```

We see that missing values are present in DiscountPrice (in Rs) and Discount Offer So, we can work on only two columns which have missing values therefore we know that the relation between DiscountPrice (in Rs), OriginalPrice (in Rs) and DiscountOffer.

We can replace the missing values in Microsoft Excel by using formula based on relation between the terms DiscountPrice (in Rs), OriginalPrice (in Rs) and DiscountOffer.

**Imputing data that is missing:** By using the following formula we can replace the missing values

$$\text{Discount price (in Rs)} = \frac{Original\ Price * (100 - Discount\ \ (in\ \%))}{100}$$

Discount (in %) = 0 then Discount price (in Rs) = 0

After imputation of missing values, we can see that there is no missing value present in the dataset

```
data.isnull().sum()
```

```
Product_id                0
BrandName                 0
Category                  0
Individual_category       0
category_by_Gender        0
OriginalPrice (in Rs)     0
Discount (in Rs)          0
Discount (in %)           0
Ratings                   0
Reviews                   0
SizeOption                0
Description               0
dtype: int64
```

## 4.4. Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of examining and analysing a dataset to understand its characteristics, identify patterns, and discover relationships between variables. It is a process that involves collecting, cleaning, transforming, and visualizing data. EDA is a process used for visualisation and summarisation of data and finding meaningful insights from the data.

Data scientists can make sure their findings are reliable and relevant to any intended business objectives by using exploratory analysis. By verifying that stakeholders are posing pertinent questions, eda further assists them. Questions concerning confidence intervals, categorical variables, and standard deviations can all be addressed by eda. After eda is finished and conclusions are made, its features can be applied to more complex data analysis or modelling, such as machine learning. EDA is an essential part of the data science process, and it can help us to get the most out of our data. By understanding the data, we can make better decisions and develop more accurate models.

Some of the most common data science tools used to perform an EDA include:

- **Python**:
  A dynamically semantic, object-oriented programming language interpreter. Its dynamic typing and dynamic binding, along with its high-level built-in data structures, make it an appealing language for quickly developing applications and for use as a scripting or glue language to join disparate components. It's crucial to determine how to handle missing values for machine learning by using Python and EDA together to find missing values in a data set.

- **R-Software:**
  A free software environment and open-source programming language for statistical computing and graphics that are backed by the R foundation. When creating statistical observations and conducting data analysis, statisticians in the field of data science frequently utilize the R language.

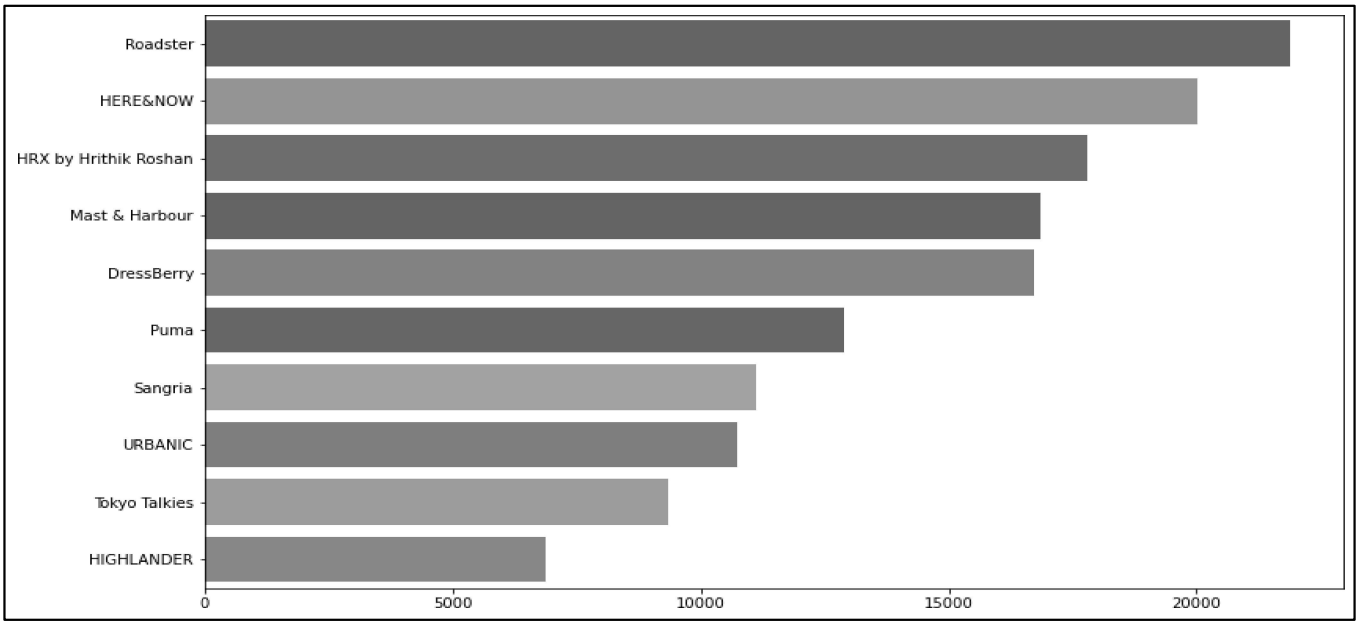
- **Tableau**:

  Tableau is a data visualization tool that is mainly used for decision-making in the business intelligence domain. Tableau facilitates data analysis and dashboard creation for insights. Tableau is a data visualization tool that is mainly used for decision-making in the business intelligence domain. Tableau facilitates data analysis and dashboard creation for insights.

- **Power BI**:

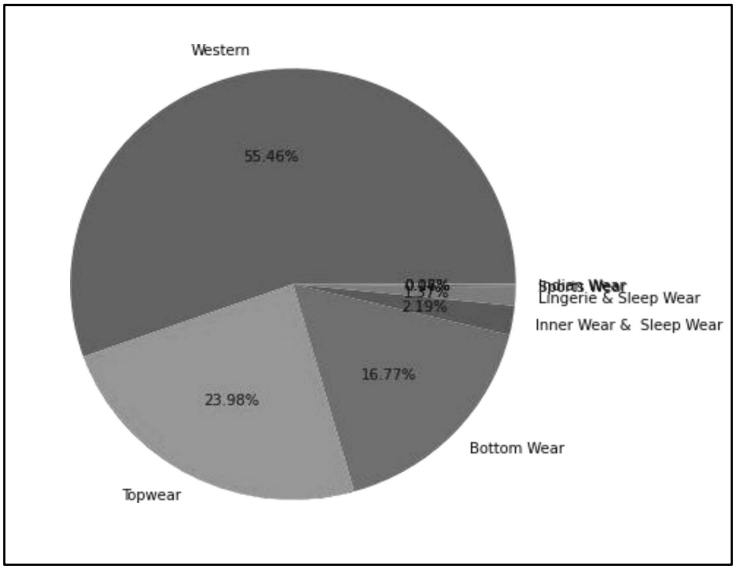
  Data analysts are the primary users of power bi, a self-service bi tool that makes data analytics accessible to staff members. Management and department representatives use Microsoft Power Bi to generate reports and forecasts that help sales and marketing representatives and give management information on how the department or specific employees are doing in terms of reaching their objectives.

1. **Sales Of Brands Based on Ratings:**



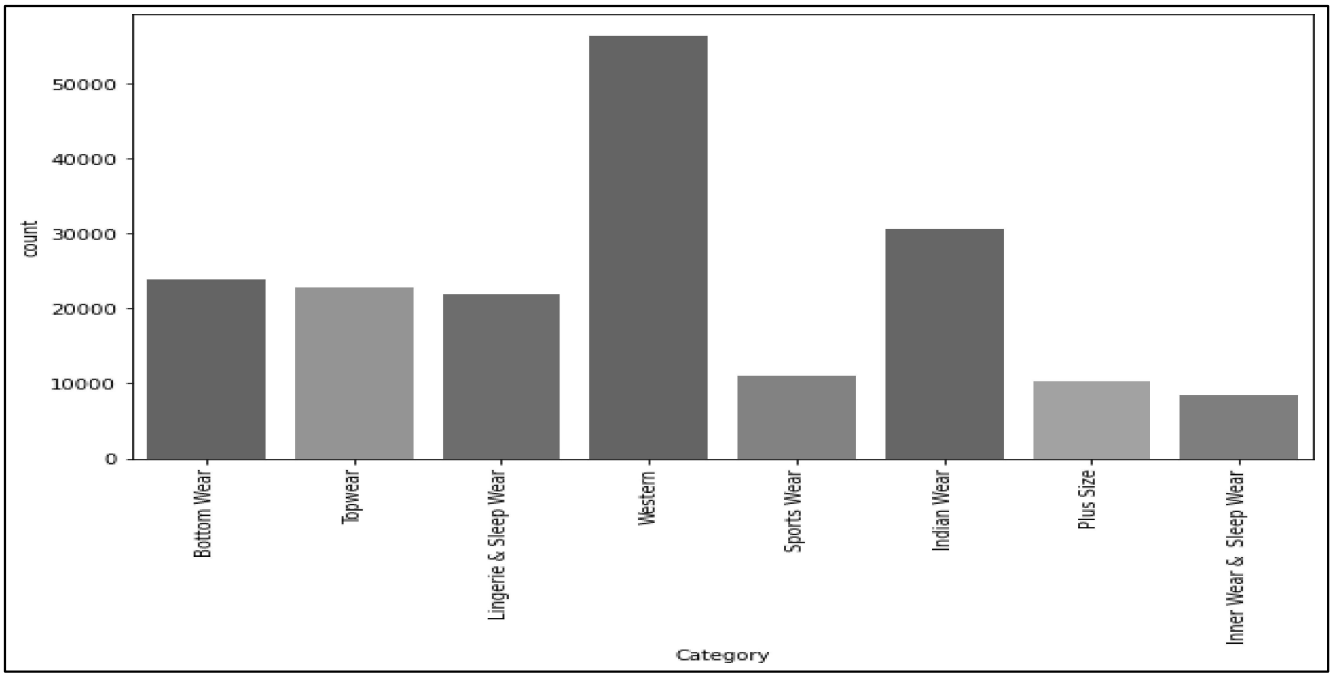

**Interpretation**: - From the above bar diagram, we can easily understand that 'Roadster' is the brand which has highest selling compare to other brands.

2. **Plot Of Category Under the BrandName (Roadster):**



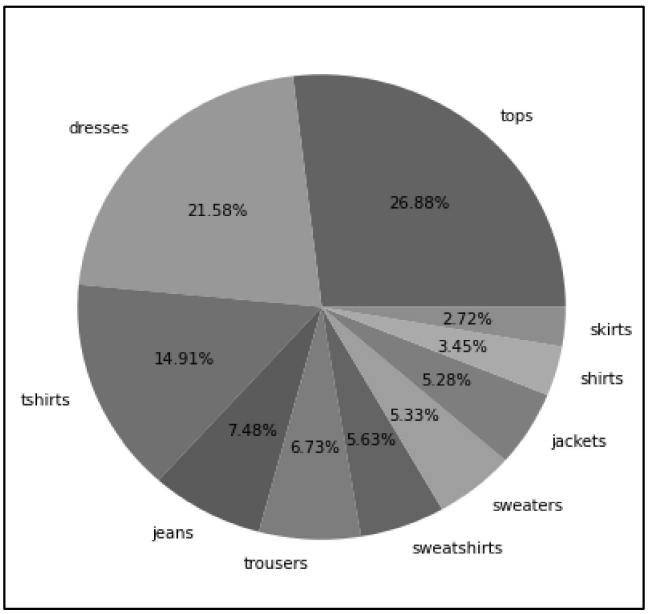

**Interpretation**: - From the above Pie chart, we can easily understand that Top brand (Roadster) under Category is 'Western (55.46%)'.

3. **Sales of Category Based on Count:**



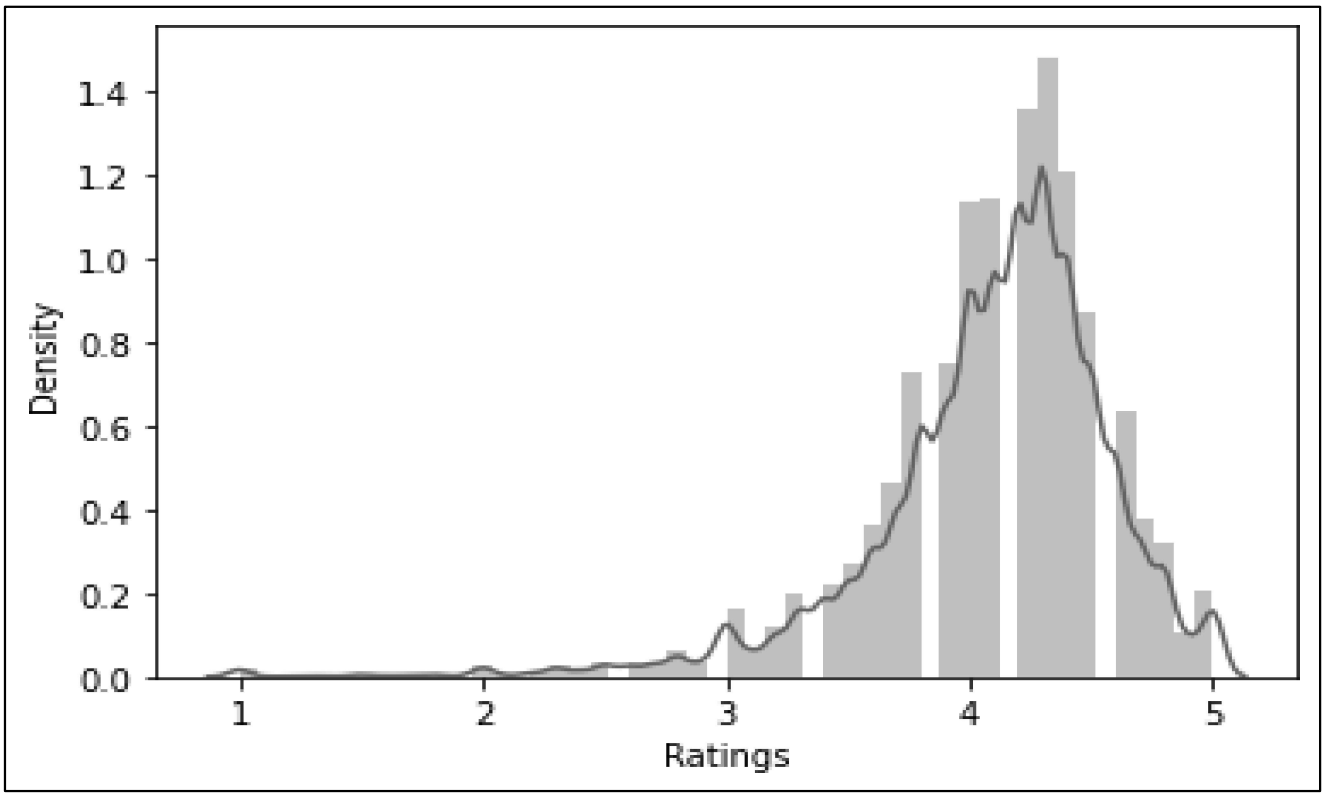

**Interpretation**: - From the above bar graph, we can visualize that 'Western' is the category which is more preferable as compare to another category.

4. **Plot of Individual Category Under the Category (Western):**




**Interpretation**: - From the above graph we can easily understand that Category (Western) under the Individual Category is 'tops (26.88%)'.

5. **Distribution Plot on Product Ratings:**




**Interpretation: -**

The distribution of ratings in the dataset is skewed to the right, with a mode of 4 and a mean of 3.7. This indicates that there are more ratings to 4 than any other rating value and also, we can easily understand from this distribution plot that most of the customer satisfy with products had positive experience and there are few customers had negative experience.

# Chapter-V

## Analysis Using Software

### 5.1. Natural Language Processing:

The ability of a computer program to comprehend spoken and written human language is known as natural language processing, or nlp. It's part of the artificial intelligence (ai) system. With its origins in the study of languages, nlp has been around for more than 50 years. It has numerous practical uses in a range of industries, such as business intelligence, search engines, and medical research.

- **What is the process of natural language processing?**

Computers can comprehend natural language just like people thanks to nlp. Natural language processing uses artificial intelligence to process and interpret real-world input spoken or written in a way that a computer can comprehend. Computers have programs to read and microphones to record audio, just as humans have various sensors like ears to hear and eyes to see. Computers have programs to process the inputs that they receive, just as humans have brains to process information. The input is eventually changed into computer-understandable code during processing.

The two primary stages of natural language processing are as follows:

1. Data preprocessing.

2. Algorithm development.

'Preparing' and 'cleaning' text data so that computers can analyze it is known as data preprocessing. Preprocessing identifies textual elements that an algorithm can use and puts data into a format that is practical.

This can be accomplished in a number of ways, including:

- Tokenization **:** This is when text is broken down into smaller units to work with.
- Stop word removal: This is when common words are removed from text so unique words that offer the most information about the text remain.
- Lemmatization and stemming**:** This is when words are reduced to their root forms to process.
- Part-of-speech tagging: this is the marking of words according to the part of speech they belong to, such as adjectives, verbs, and nouns.

An algorithm is created to process the data after it has undergone preprocessing. Although there are numerous natural language processing algorithms, but two main types are commonly used:

- Rules-based system**:** This system makes use of meticulously crafted linguistic rules. This method is still in use today, having been employed early in the development of natural language processing.

- Machine learning-based system: Statistical techniques are used by machine learning algorithms. They are fed training data to help them learn how to perform tasks, and as more data is processed, they modify their approaches accordingly. Natural language processing algorithms use a combination of neural networks, deep learning, and machine learning to refine their own rules via repeated processing and learning.

## Extract keywords from product descriptions by using NLP technique:

Extracting keywords from product descriptions using Natural Language Processing (NLP) techniques is a valuable approach for enhancing product search and recommendation systems. By identifying the key terms and phrases that accurately represent the products, NLP algorithms can effectively classify and categorize items, enabling users to discover relevant products more efficiently.

Firstly, we are applied Term Frequency-Inverse Document Frequency (TF-IDF) score NLP techniques to our dataset for extraction of keywords from the product description.

TF-IDF measures the importance of a term within a document relative to its overall presence in the corpus (collection of documents). It assigns higher weights to terms that appear frequently in a specific document while penalizing those that are common across all documents.

Here is a general an overview of the process:

1. Tokenization: Split each product description into individual words or tokens.
2. Remove Stopwords: Remove common words that don't carry much meaning (e.g., "the," "and," "is").
3. Calculate Term Frequency (TF) : Count the number of times each word appears in a product description.
4. Calculate IDF (Inverse Document Frequency): IDF is a measure of how important a word is across all product descriptions. It's calculated as the logarithm of the total number of descriptions divided by the number of descriptions containing the word.
5. Calculate TF-IDF: Multiply TF by IDF for each word in each product description.
6. Select Keywords: Choose the top TF-IDF scored words as keywords for each product description.

Here is an overview simplified example in Python using the scikit-learn library:

**Output:**

| | Description | Keywords | tfidf_Scores |
|---|---|---|---|
| 6040 | wrogn men navy blue striped henley neck t shirt | men | 2.776652 |
| 23789 | orchid blues navy blue sweetheart neck net reg... | blue | 1.710793 |
| 135930 | arise men olive green sweatshirt | navy | 1.605198 |
| 56364 | laado pamper yourself women maroon hem desig... | women | 1.147269 |
| 21339 | vimal charcoal grey lounge shorts d13 | green | 1.794956 |
| 85627 | spykar men black white tapered fit striped re... | shorts | 1.960197 |
| 134151 | tommy hilfiger men navy blue off white slim f... | grey | 1.796764 |
| 89062 | bannos swagger yellow printed maxi nightdress | fit | 1.556638 |
| 59042 | roadster men mustard brown solid pure cotton r... | solid | 3.562304 |
| 177804 | sg leman men navy blue angrakkha kurta | cotton | 2.389098 |
| 67473 | anouk women maroon solid a line dress with pri... | design | 1.375201 |
| 108957 | dixcy scott men blue colourblocked sports shorts | maroon | 1.346159 |
| 70081 | klamotten pink self design baby doll yy07 | top | 1.428697 |
| 81636 | reebok men green polo collar speedwick t shirt | regular | 1.319739 |
| 171749 | kalini green white pure cotton checked dharma... | white | 1.815146 |
| 54946 | cherokee women grey solid pure cotton top | pink | 1.752160 |
| 90433 | roadster men charcoal grey skinny fit light fa... | pure | 1.543760 |
| 138841 | boston club women black purple colourblocked ... | shirt | 0.981267 |
| 87235 | marks spencer pink solid extended sleeves top | dress | 1.395808 |
| 104728 | biba navy blue maroon a line midi dress | trousers | 2.331103 |

**Interpretation:**

From the above output we can conclude that NLP is technique used to extract keywords from text by identifying the most important and relevant words and phrases based on Term Frequency-Inverse Document Frequency (TF-IDF) scores.

## Cluster product descriptions and Individual category into similar groups by using NLP technique(K-Means):

K-Means clustering is a popular unsupervised machine learning algorithm that can be used to group data into a predefined number of clusters. The algorithm works by iteratively assigning data points to the nearest cluster centroid, and then updating the centroid to be the mean of the assigned data points.

Firstly, we are applied K-means clustering with NLP techniques to our dataset for cluster product descriptions and individual categories into similar groups you will need to follow these steps:

1. Data Preprocessing:
    - Cleaning: Remove stop words, punctuation, and special characters from product descriptions.
    - Normalization: Lemmatize or stem words to their base forms.
    - Vectorization: Convert text descriptions into numerical vectors using techniques like TF-IDF.
2. Feature Engineering:
    - Extract additional features: Analyze product descriptions to extract relevant features like Colour, material, size, etc.
    - Combine features: Combine text-based features with category information to create multidimensional vectors representing each product.
3. K-Means Clustering:
    - Choose the number of clusters: Use methods such as the elbow method or silhouette score to find the ideal number of clusters.
    - Initialize centroids: Randomly initialize cluster centroids or use seed vectors based on existing categories.
    - Iterate: Assign each product to the nearest cluster based on distance metrics like Euclidean distance.

Here is an example of how to cluster product descriptions using K-Means in Python:

**Cluster 1:**

| id | Individual category | Description |
|---|---|---|
| 12 | kurta-sets | anubhutee women pink white printed kurta with trousers dupatta |
| 16 | kurtas | vishudh women blue pink floral print a line kurta |
| 104 | kurtas | moda rapido women white blue printed straight kurta |
| 17 | kurta-sets | sangria women green off white printed kurta with palazzos dupatta |
| 81 | tops | herenow black ethnic motifs a line top |
| 183 | palazzos | clora creation women white embroidered straight palazzos |
| 28 | kurtas | sangria women black green printed straight kurta |
| 237 | kurta-sets | varanga calm blue and grey yoke design kurta set |
| 60 | kurtas | anayna women beige pink screen print straight kurta |
| 259 | tops | sangria blue ethnic motifs a line pure cotton top |
| 62 | kurta-sets | sangria women fuchsia white layered printed kurta with trousers |

**Cluster 2:**

| id | Individual_category | Description |
|---|---|---|
| 1 | track-pants | locomotive men black white solid slim fit track pants |
| 10 | tights | hrx by hrithik roshan men rapid dry training tights |
| 93 | tshirts | hrx by hrithik roshan men navy blue solid training t shirt |
| 39 | track-pants | hrx by hrithik roshan men olive green camo athleisure track pants |
| 96 | tights | hrx by hrithik roshan women navy solid rapid dry training tights |
| 47 | socks | hrx by hrithik roshan men quarter length pack of 3 terry socks |
| 136 | jackets | hrx by hrithik roshan men black grey camouflage printed running bomber jacket |
| 180 | bra | hrx by hrithik roshan seamless black grey lightly padded rapid dry sports bra 3376 |
| 167 | tights | hrx by hrithik roshan women navy blue yoga seamless solid tights |
| 193 | lounge-pants | vimal men pack of 2 solid cotton lounge pants |

**Cluster 3:**

| id | Individual_category | Description |
|---|---|---|
| 7 | tops | mayra pink embroidered a line pure cotton top |
| 35 | bath-robe | elevanto women fuchsia pink solid bath robe |
| 50 | tops | harpa women pink printed wrap top |
| 65 | dresses | sangria teal blue coral pink floral print a line dress |
| 131 | night-suits | claura women yellow white printed night suit cot 108 yellow lt |
| 133 | tights | neu look fashion women black pink colourblocked gym tights |
| 214 | night-suits | etc women navy blue pink floral print night suit |
| 265 | sarees | saree mall women pink solid saree |
| 154 | dresses | sera women burgundy pink printed tailored dress |
| 293 | dresses | uf black pink floral print maxi dress |

**Cluster 4:**

| id | Individual_category | Description |
|---|---|---|
| 18 | trousers | tokyo talkies women olive green regular fit solid cargos |
| 30 | tshirts | roadster women green teal green pack of 2 printed round neck t shirts |
| 31 | dresses | vishudh women navy blue green maxi dress |
| 32 | jumpsuit | cottinfab olive green solid cold shoulder jumpsuit |
| 42 | trousers | sassafras women olive green peg trousers |
| 45 | tops | mayra women mustard yellow green ditsy floral printed cinched waist top |
| 48 | tshirts | moda rapido men olive green yellow striped cotton pure cotton t shirt |
| 69 | kurta-sets | anouk women green printed kurti with dhoti pants |
| 221 | dresses | sassafras off white green floral printed a line dress |
| 182 | trousers | highlander men olive green slim fit solid joggers |

**Cluster 5:**

| id | Individual_category | Description |
|---|---|---|
| 6 | trousers | highlander men olive green slim fit solid regular trousers |
| 166 | sweatshirts | hm men grey relaxed fit sweatshirt |
| 229 | tshirts | highlander men charcoal grey slim fit solid henley neck t shirt |
| 27 | shorts | wisstler women navy pink chevron print regular fit shorts |
| 539 | track-pants | maniac men black solid slim fit joggers |
| 793 | shorts | highlander men navy blue striped slim fit regular shorts |
| 2128 | blazers | sassafras women black solid slim fit single breasted formal pure cotton blazer |
| 151 | trousers | roadster men khaki slim fit solid chinos |
| 2627 | leggings | rangmanch by pantaloons women red solid slim fit leggings |
| 50621 | sweatshirts | puma men grey melange solid evostripe full zip slim fit hooded sweatshirt |

**Interpretation:**

From the above formed clusters, we can conclude that K-means algorithm can be used to cluster product descriptions and individual category into similar groups and similar corpus and by using this clusters we can easily find out the category that actually we want based on the product description.

## 5.2. Recommendation System:

A recommendation system is a technology that suggests items, products, or content to users based on their preferences, behaviour, or characteristics. It's like having a personalized assistant that understands your likes and dislikes and helps you discover new things that you might enjoy.

In the case of the Myntra fashion product dataset, the recommendation system would analyse various factors such as user browsing history, purchase behaviour, product details like brand, style, colour and even user feedback or ratings. This information helps generate better recommendations for each customer.

A recommendation system can be broadly categorized into two types:

- Content-based filtering: This type of algorithm recommends items that are similar to items that the user has liked or purchased in the past. For example, if a user has purchased a lot of books about science, a content-based filtering algorithm might recommend other books about science.
- Collaborative filtering: This type of algorithm recommends items that are popular with similar users. For example, if a user has similar taste in movies to another user, a collaborative filtering algorithm might recommend movies that the other user has liked.
- Hybrid filtering: This type of algorithm combines these two methods or it is the combination of content-based filtering and collaborative filtering.

Recommendation systems are a powerful tool that can be used by businesses to improve sales, customer satisfaction, and user engagement. As the amount of data available about users continues to grow, recommendation systems are likely to become even more sophisticated and effective.

## To build a product recommendation system based on Price, Ratings, Reviews and Individual Category:

To Build a product recommendation system based on Price, Ratings, Reviews and Individual Category involves creating a model that suggests products to users based weighted average score and also several steps, including data preparation, model development, and evaluation.

Here's a simplified example using Python with some commonly used libraries:

**Input:**

```
# Example usage
Ratings = 4
Reviews = 500
Price = 1000
Individual_category = "tops"  # Replace with the desired category

recommended_products = recommend_products(Ratings, Reviews, Price, Individual_category)

# Print the recommended products
print(recommended_products)
```

**Output:**

| | Product_id | BrandName | Category | Individual_category |
|---|---|---|---|---|
| 5 | 2490950 | Mast & Harbour | Western | tops |
| 414 | 11373682 | Roadster | Western | tops |
| 1585 | 8330195 | Roadster | Western | tops |
| 730 | 2490953 | Mast & Harbour | Western | tops |
| 1187 | 11373642 | Roadster | Western | tops |
| 380 | 11735666 | her by invictus | Western | tops |
| 1336 | 2490952 | Mast & Harbour | Western | tops |
| 713 | 807931 | La Zoire | Western | tops |
| 276 | 12009564 | Tokyo Talkies | Western | tops |
| 1628 | 10964506 | Roadster | Western | tops |

| category_by_Gender | OriginalPrice (in Rs) |
|---|---|
| Women | 599.0 |
| Women | 599.0 |
| Women | 499.0 |
| Women | 599.0 |
| Women | 599.0 |
| Women | 699.0 |
| Women | 599.0 |
| Women | 699.0 |
| Women | 799.0 |
| Women | 599.0 |

| Discount (in %) | Ratings | Reviews | SizeOption |
|---|---|---|---|
| 40 | 4.4 | 999 | XS, S, M, L, XL |
| 55 | 4.2 | 918 | XS, S, M, L, XL |
| 40 | 4.2 | 749 | XS, S, M, L, XL |
| 45 | 4.2 | 866 | XS, S, M, L, XL |
| 60 | 4.5 | 798 | XS, S, M, L, XL |
| 50 | 4.3 | 925 | S, M, L, XL, XXL |
| 40 | 4.3 | 782 | XS, S, M, L, XL |
| 20 | 4.3 | 868 | S, M, L, XL |
| 47 | 4.4 | 946 | S, M, L, XL |
| 45 | 4.2 | 742 | XS, S, M, L, XL |

| Description | weighted_score |
|---|---|
| mast harbour women yellow solid tank top | 7.338230 |
| roadster women white black solid a line pure ... | 6.436728 |
| roadster women mustard yellow styled back pure... | 6.304208 |
| mast harbour women black solid top | 6.072120 |
| roadster women black white solid a line pure ... | 5.994992 |
| her by invictus navy blue white floral pleate... | 5.690272 |
| mast harbour women pink solid tank top | 5.613689 |
| la zoire green sheer knotted top | 5.339628 |
| tokyo talkies navy blue white vertical stripe... | 5.209512 |
| the roadster lifestyle co women mustard yellow... | 5.202671 |

## 5.3. Sentiment Analysis:

Sentiment analysis is a technique used to determine the sentiment or emotion expressed in a piece of text, such as a review, comment, or social media post. It involves analysing the words, phrases, and context of the text to understand whether the sentiment is positive, negative, or neutral. This analysis is particularly useful for extracting insights from large amounts of textual data, such as social media posts, customer reviews and ratings or news articles.
Sentiment analysis, also known as opinion mining or emotional AI, is a field of natural language processing (NLP) that deals with the identification, extraction, analysis, and understanding the information in text or speech. It aims to extract and quantify the emotional tone of a piece of text, identifying specific emotions (happy, sad, angry, etc.).

There are a number of different methods for sentiment analysis. Some of the most common methods are:

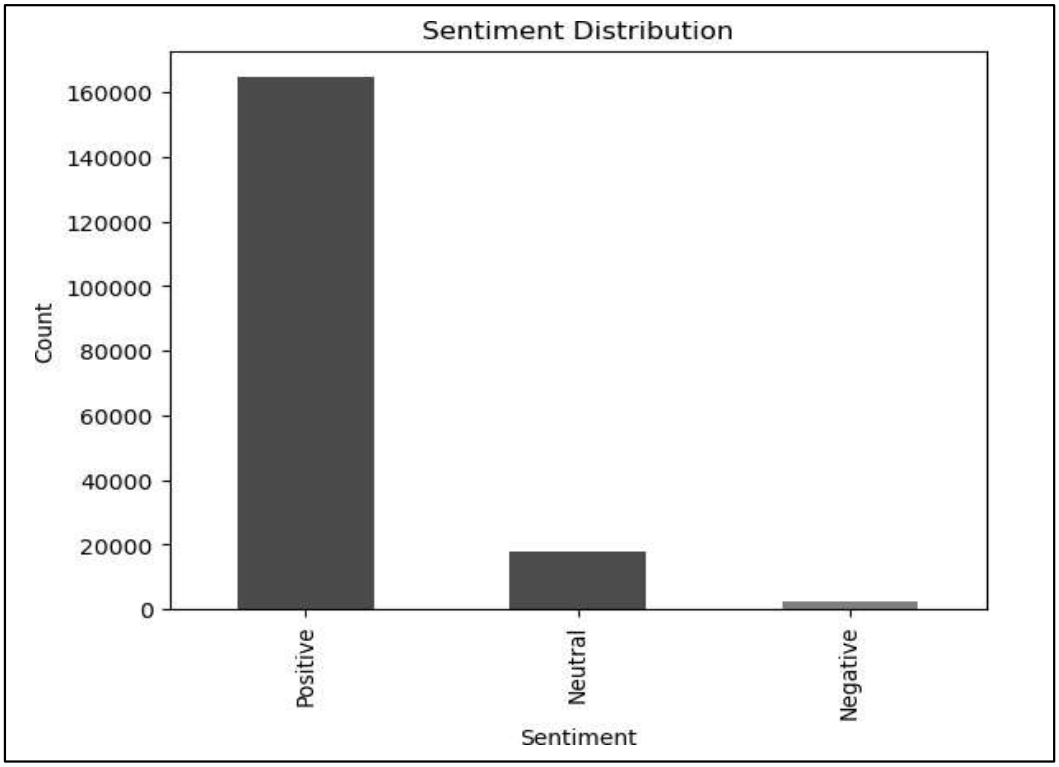
- Lexicon-based methods: These techniques employ a predetermined list of terms and expressions that are connected to either a positive, negative, or neutral emotion. The sentiment of a text is determined by the frequency of these words and phrases.
- Machine learning methods: These methods use machine learning algorithms to learn the relationship between words and sentiment. The algorithms are trained on a large amount of labeled text, and then they are used to predict the sentiment of new text.
- Hybrid methods: These methods combine lexicon-based methods and machine learning methods. This may contribute to raising the sentiment analysis's accuracy.
- Sentiment analysis has become increasingly sophisticated with the advancement of NLP techniques and the availability of large labeled datasets. It is now widely used in various industries to gain valuable insights from the large amount of unstructured text data generated daily.

## To develop a multi-class sentiment classification model:

Firstly, we are performing sentiment classification on product ratings then calculate the average sentiment score for each rating and we are distributing the product rating column into three category such as positive, negative and neutral as shown in the following table:

| Product Rating | Sentiment Classification |
|---|---|
| 1.0 to 2.5 | Negative Sentiment |
| 2.6 to 3.5 | Neutral Sentiment |
| 3.6 to 5.0 | Positive Sentiment |

Based on this sentiment classification we are create new column sentiment classification in our dataset and draw the bar diagram for visualisation of sentiments and total count of sentiments.



**Interpretation:** from the above bar diagram of sentiment distribution, we can easily analyse that the count of positive sentiment is highest as compare to neutral and negative sentiment.

Here we are calculating the count of each sentiment category in the column

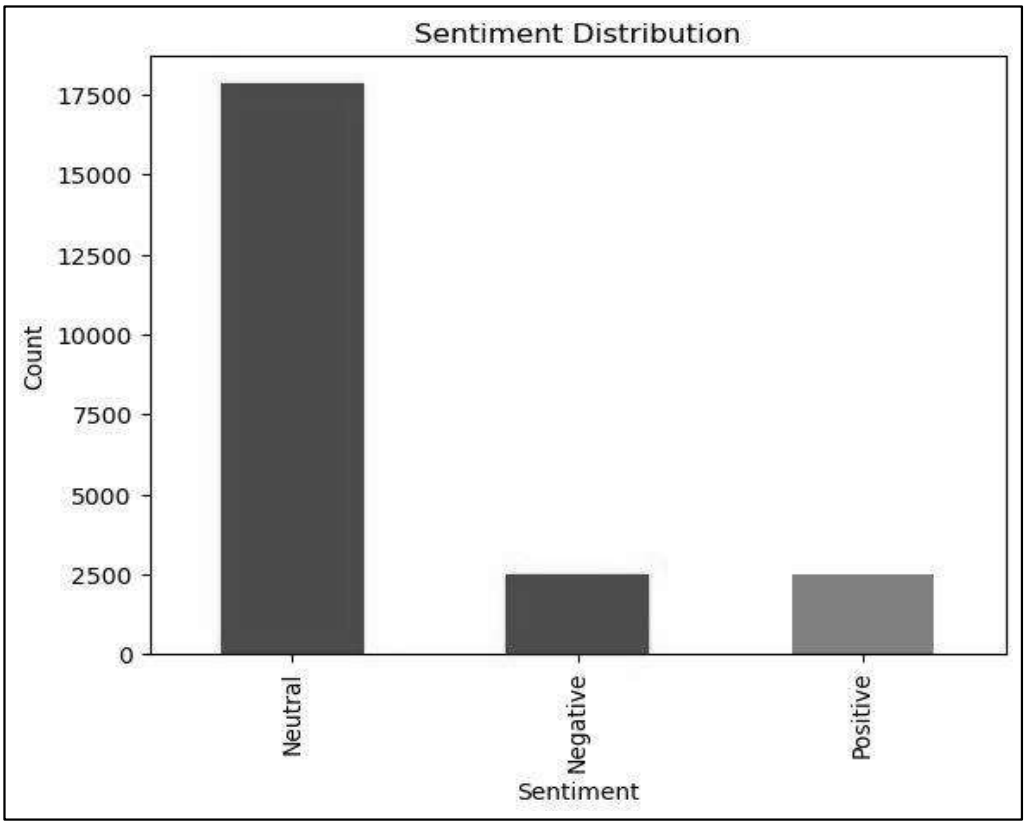
```
Sentiment_Category
Positive    164599
Neutral      17837
Negative      2477
Name: count, dtype: int64
```

So, by observing the count of sentiments we can say that our data is imbalance so we have to handle the imbalance dataset by using the technique such as

Sampling Techniques:

1. **Under-sampling**: To match the size of the minority class, this method entails reducing the number of data points from the majority class. Random under-sampling is a simple method that randomly removes instances from the majority class. However, it may discard valuable information and reduce the overall size of the training set.
2. **Over-sampling**: This approach involves replicating data points from the minority class to increase its representation in the training set. Oversampling at random only results in duplicates of the minority class. However, it can lead to overfitting and introduce artificial correlations.
3. **Synthetic Minority Oversampling Technique (SMOTE):** SMOTE is an effective oversampling technique that creates new synthetic minority class instances by interpolating between existing minority class data points. This approach helps preserve the inherent distribution of the original data while increasing the size of the minority class
4. **Use ensemble methods**: such as bagging and boosting, with algorithms that inherently handle imbalanced data well. Algorithms like Random Forest and AdaBoost often perform better on imbalanced datasets.

For our dataset we are implement the Random Under-Sampling technique to handle the imbalance sentiment class present in the data. So, by using the under-sampling method it reduces number of positive sentiments up to number of negative sentiments and neutral sentiment class remain same as shown in above bar graph;




**Interpretation:** from the above bar diagram of sentiment distribution, we can easily analyse that the count of positive sentiment and negative sentiment are same & neutral sentiment remains as it is.

Here we are calculating the count of each sentiment category in the column:

```
Sentiment_Category
Neutral     17837
Negative     2477
Positive     2477
Name: count, dtype: int64
```

After handling the imbalance data then we have to develop a multi-class sentiment classification model that classifies product ratings as positive, negative, or neutral.

## ❖ Sentiment Analysis Using Multinomial Naive Bayes:

Multinomial Naive Bayes (MNB) is a simple and effective machine learning algorithm commonly used for sentiment analysis. It is based on Bayes' theorem, which provides a framework for calculating probabilities based on conditional statements. In the context of sentiment analysis, MNB is used to classify text into predefined sentiment categories, such as positive, negative, or neutral.

The MNB algorithm assumes that the presence or absence of words in a document is independent of other words, given the sentiment category. However, MNB often performs well in practice, especially for tasks like sentiment analysis where the vocabulary size is large and the assumption of independence is less restrictive.

Here's a simplified explanation of how MNB works for sentiment analysis:

1) Data Preprocessing: Clean and prepare the text data by removing irrelevant characters, applying stemming or lemmatization, and converting text to lowercase.

2) Feature Extraction: Represent the text data as a feature vector. Another option is to use TF-IDF (Term Frequency-Inverse Document Frequency), which takes into account not only the frequency of a word in a document but also its importance in the entire dataset

3) Training: the algorithm calculates the probability of it belonging to each class based on the features (words) present in the document. The class with the highest probability is predicted as the sentiment of the document.

4) Model Evaluation: The performance of the model is evaluated using a test dataset. Common metrics for sentiment analysis include accuracy, precision, recall, and F1-score.

Here's an example of how to interpret the performance of a Multinomial Naive Bayes algorithm for sentiment analysis on dataset using Python:

**Output:**

| Accuracy: | 0.7911822768 |
|---|---|

**Classification Report:**

| Column1 | Precision | recall | f1-score | Support |
|---|---|---|---|---|
| Negative | 0.12 | 0 | 0 | 464 |
| Neutral | 0.79 | 1 | 0.88 | 3612 |
| Positive | 0.38 | 0.01 | 0.01 | 483 |
| | | | | |
| Accuracy | | | 0.79 | 4559 |
| macro avg | 0.43 | 0.34 | 0.3 | 4559 |
| weighted avg | 0.68 | 0.79 | 0.7 | 4559 |

**Model Evaluation:**

To assess the performance of the trained MNB classifier on the testing set, we calculated the following metrics:

• Accuracy: The proportion of correctly classified reviews

$$= \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

• Precision: The proportion of positive predictions that are actually correct

$$= \frac{TP}{(TP+FP)}$$

• Recall: The proportion of actual positive reviews that are correctly identified

$$= \frac{TP}{(TP+FN)}$$

• F1-score: The harmonic mean of precision and recall

**Result and Conclusion:**

- **Overall Accuracy :** The overall accuracy of the MNB classifier is 0.791, indicating that it correctly classified 79.1% of the product ratings. This is a relatively high accuracy, suggesting that the classification is able to effectively distinguishing between positive, negative, and neutral sentiments.

- **Sentiment Class Performance:**

1. Negative Reviews: The classifier has a precision of 0.12 for negative reviews, meaning that only 12% of the reviews it classified as negative were actually negative. This suggests that the classification tends to over-predict negative sentiment.

2. Neutral Reviews: The classifier has a precision of 0.79 and a recall of 1 for neutral reviews, indicating that it accurately identified neutral reviews with a high degree of confidence.

3. Positive Reviews: The classifier has a precision of 0.38 and a recall of 0.01 for positive reviews. This suggests that the classifier tends to under-predict positive sentiment.

## Result and Discussion

1) We can determine which brands and clothing accessories in the dataset are more popular among customers based on sales of category and individual category, ratings, reviews, and price by using exploratory data analysis.

2) We can understand from distribution plot product ratings that most of the customer satisfy with products had positive experience and there are few customers had negative experience.

3) We can conclude that NLP is technique used to extract keywords from text by identifying the most important and relevant words and phrases based on TF-IDF scores. By capturing keywords from product descriptions NLP technique can improve marketing strategies, e-commerce experiences, and search capabilities. It helps in identifying market trends, and improving product quality.

4) We can conclude that K-means clustering improves product classification and user experience by automatically grouping similar products into clusters. This enhances search results, accuracy and relevancy, reducing time spent searching and enhancing customer satisfaction.

5) We can build a recommendation system based on rating, review, individual category and price can improve customers purchasing experiences. The precise and customized recommendations makes decision-making easier using this method customers find goods within their financial limits thus recommendation system is accurate and relevant to the needs of customers.

6) We can conclude that the overall accuracy of the Multinomial Naive Bayes classifier is 0.791, indicating that it correctly classified 79.1% of the product ratings. This is a relatively high accuracy, suggesting that the classification is able to effectively distinguishing between positive, negative, and neutral sentiments.

# Limitation and Future Scope

## Limitation:

1) The dataset only includes products from the Myntra website, which means that it does not cover the entire fashion market. This could limit the generalizability of any models trained on the dataset. the limitation that you could mention in your project report is the potential bias in the dataset.
2) the dataset is collected from Myntra, it might not represent the entire fashion industry or capture the diversity of fashion trends and styles.
3) One of the main limitations of the Myntra Fashion Product Dataset is the data quality. The dataset is scraped from the Myntra website, which means that it may contain errors and inconsistencies. For example, there may be missing values, incorrect product names or descriptions, and duplicate entries.

## Future Scope:

1) The future scope of the Myntra Fashion Product Dataset is limited only by the creativity of the researchers and developers who use it. As the dataset grows and becomes more comprehensive, it will become an even more valuable resource for the fashion industry.
2) A fashion designer could use the dataset to identify trends in fashion and to predict future trends. This could help the designer to create products that are likely to be popular with consumers.
3) The future scope of the Myntra fashion product dataset for your project report, there are a few exciting possibilities. One idea could be to explore the temporal aspect of the dataset and analyze how fashion trends evolve over time.
4) Additionally, you could explore the application of advanced machine learning techniques, such as deep learning or generative models, to enhance the dataset and generate more accurate fashion recommendations.

# References


1. Jaiswal, S., Patel, D., Vempati, S., & Saiswaroop, K. (2023). Product Review Image Ranking for Fashion E-commerce. arXiv preprint arXiv:2308.05390.

2. Luciano Rivera, G. (2023). Data-driven clustering for new garment forecasting (Doctoral dissertation, Massachusetts Institute of Technology).

3. Sharma, A. K., Bajpai, B., Adhvaryu, R., Pankajkumar, S. D., Gordhanbhai, P. P.& Kumar, A. (2023). An efficient approach of product recommendation system using NLP technique. Materials Today: Proceedings, 80, 3730-3743.

4. Umaashankar, V., & Prakash, A. (2019). Atlas: a dataset and benchmark for e-commerce clothing product categorization. arXiv preprint arXiv:1908.08984

5. https://www.kaggle.com/datasets/manishmathias/myntra-fashion-dataset

6. https://newspatrolling.com/myntra_company_profile/

7. https://en.m.wikipedia.org/wiki/amazon-(company)

8. https://www.myshared.ru/amp/139990

9. http://www.thehindu.com/business/industry/govt-permits-100-per-cent-fdiin-online-market-places/article8409495.ece.