Insights and Recommendations from Reviews

- Aishwarya Nair

1. Introduction

In the digital age, customer reviews have become a critical component for understanding consumer sentiment and driving business decisions. Online retailers, in particular, rely heavily on customer feedback to refine their products, enhance customer satisfaction, and maintain a competitive edge. This report delves into the analysis of customer reviews for a generic online retailer, aiming to uncover insights and trends.

The objective here is to identify the key factors that influence customer satisfaction, pinpoint recurring pain points, and highlight aspects that customers appreciate. By leveraging data processing and visualization tools along with advanced Natural Language Processing techniques, unstructured review text can be transformed into actionable insights. This comprehensive analysis will cover several critical areas:

- Identifying patterns or trends that stand out in different segments of the data.
- Obtain a statistical summary for the numerical attributes within the dataset.
- Perform overall sentiment analysis to gauge the general mood of the reviews.
- Identify common themes and extract common keywords within the reviews.
- Identify pain points from the negative reviews to isolate specific issues that customers are facing, enabling targeted improvements.
- Analysis of positive reviews to identify strengths and areas that are working well, which can be leveraged for marketing and product development.

To initiate the analysis, several key questions were identified that could be addressed using the data. These questions aim to uncover valuable insights and guide the direction in identifying pain points and areas for improvement. The further analysis based on these questions helped to identify key insights, which are discussed in detail later in the report. The analysis was done using JupyterLab.

2. Objectives

Since the given problem was open ended, the main idea was to think from the perspective of the company (retailer) and identify how the different products perform, understand the customer-product affinity, how existing products or processes can be improved, how to understand customer segments better and finally the pain areas that needs to be addressed.

With this in mind few questions were identified which could be addressed by means of this analysis. It can be categorized as follows:

1. **Product Type and Ratings**
    a. Which products are most reviewed?
    b. Which products are most popular under each type?

c. What are the most frequently reviewed product types?
d. How do their ratings compare?

2. **Age Groups, Rating and Reviews**
    a. Are there any noticeable trends in ratings and reviews based on age groups within each product type?
    b. Do certain product types receive consistent ratings across different age groups?
    c. How does age relate to changes in ratings and types of products reviewed? Does age of a person affect the rating?
    d. How does review sentiment relate to the ratings given?
    e. Are there any noticeable trends in sentiment of reviews based on age groups within each product type?
    f. What is the sentiment for each type of product? Can it be quantified with a rating? Are there any discernible patterns?
    g. What are the top reviewed products within different age groups?
3. **Product Type, Rating and Reviews**
    a. Is there a particular product or product type that consistently performs the best or the worst in terms of ratings and reviews?
    b. What are the most frequent review words for each product type?
4. **Reviews Sentiment**
    a. What is the overall sentiment of text reviews?
    b. Does the age of the customer influence the words used in the reviews?
5. **Review Analysis**
    a. Are there specific issues commonly mentioned in negative reviews as:
        i. Delivery problems (e.g., late delivery)
        ii. Damaged product
        iii. Size discrepancies
        iv. Color not matching the product description
        v. Size inclusivity
        vi. Quality of the product (stitching issues/ torn product)
    b. Are there specific keywords commonly mentioned in the positive reviews?

3. Data Dictionary

The data was studied to obtain a high-level understanding of its contents and structure. Basic statistics such as the number of rows and columns, data types, and summary statistics, were calculated to gain insights into the dataset's overall characteristics.

This table presents a structured overview of each column in the DataFrame, including the column name, description, data type, and non-null count.

Column	Description	Data Type	Non-Null Count
Product_ID	Unique identifier for each product	int64	23486

Age	Age of the reviewer	int64	23486
Title	Title of the review provided by the reviewer	object	19676
Text_Review	Detailed text review provided by the reviewer	object	22641
Rating	Rating provided by the reviewer	int64	23486
Type	Type of the product	object	23472

Sample Dataset and Dataset Size

```
Dimensions of the data: 23486 rows and 6 columns
```

	Product_ID	Age	Title	Text_Review	Rating	Type
0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	Intimate
1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	Dresses
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	Dresses
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	Bottoms
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	Tops

Data Statistics

	count	mean	std	min	25%	50%	75%	max
Product_ID	23486.0	918.118709	203.298980	0.0	861.0	936.0	1078.0	1205.0
Age	23486.0	43.198544	12.279544	18.0	34.0	41.0	52.0	99.0
Rating	23486.0	4.196032	1.110031	1.0	4.0	5.0	5.0	5.0

Checking the data for duplicates and missing values. There were 50 rows of duplicate data entries. These were identified and removed for further analysis. The data was then studied for missing and unique values within different attributes.

	Product_ID	Age	Title	Text_Review	Rating	Type
Unique_Count	1206	77	13993	22634	5	6

	Product_ID	Age	Title	Text_Review	Rating	Type
Missing_Count	0	0	3760	796	0	14

From this, it can be noted that there are 796 entries with no text reviews. Since this study focuses on evaluating the reviews, the entries with missing text reviews can be dropped for this study.

4. Exploratory Data Analysis

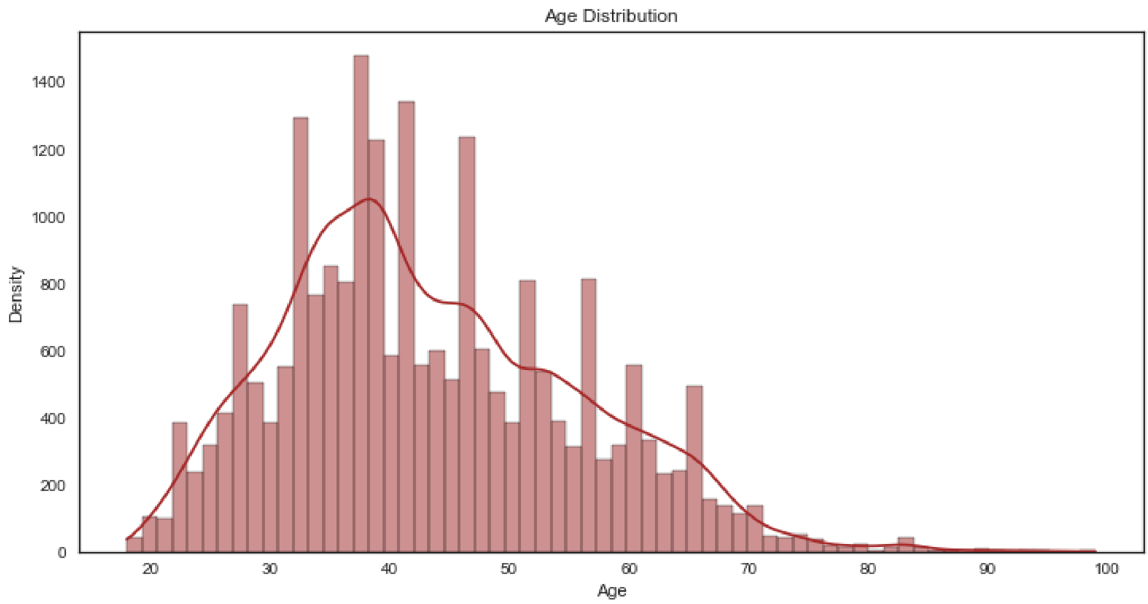
Analyzing the data with respect to the different attributes within the dataset, one at a time.

4.1 Age:

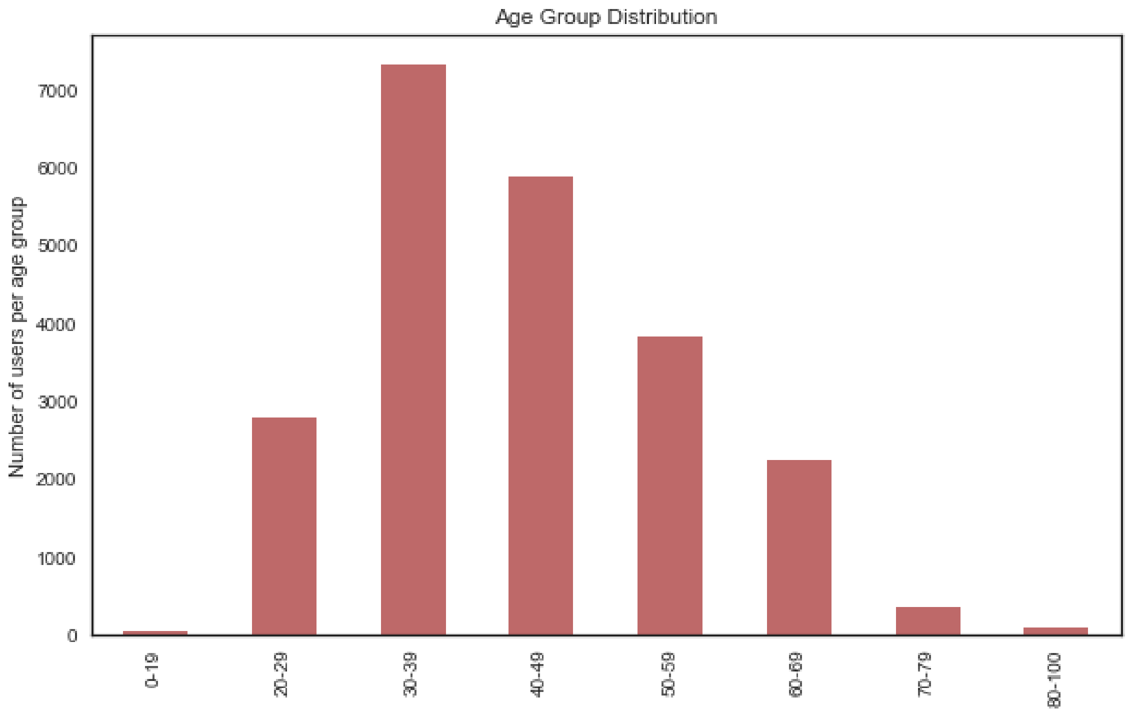
Statistics of the attribute ‘Age’ gives the following information.

- The minimum age amongst the reviewers = 18
- The maximum age amongst the reviewers = 99
- The mean age of the reviewers = 43

Analyzing the distribution of the 'Age' attribute through a histogram offers valuable insights into the demographic makeup of reviewers. By visually representing the spread of ages and the frequency of reviewers within each age group, this histogram provides an understanding of the age distribution among the dataset's reviewers.



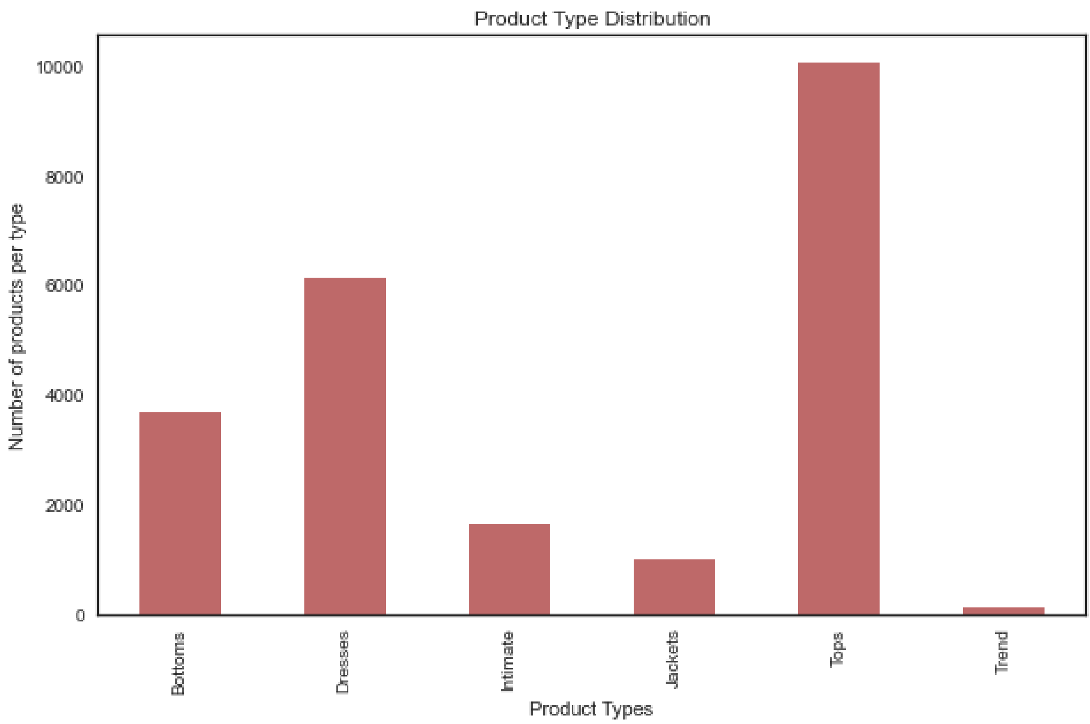
This shows that reviewers are mostly between the ages of 30 and 50. The distribution is slightly right skewed but normal.
Using the distribution as a reference, the age attribute can be divided into buckets of 10 years to conduct further analysis. The distribution of users within different age groups can be shown as follows.



This bar plot reiterates the fact that most reviewers are in the age group of 30-39, followed by those in the group of 40-49. The number of reviewers in the age groups of <19 and 80-100 are far less, comprising approximately 0.2% and 0.4% of the total, respectively. They do not effectively represent the broader demographic distribution of reviewers, indicating potential gaps in data collection or participation from these age groups.

4.2 Product Type

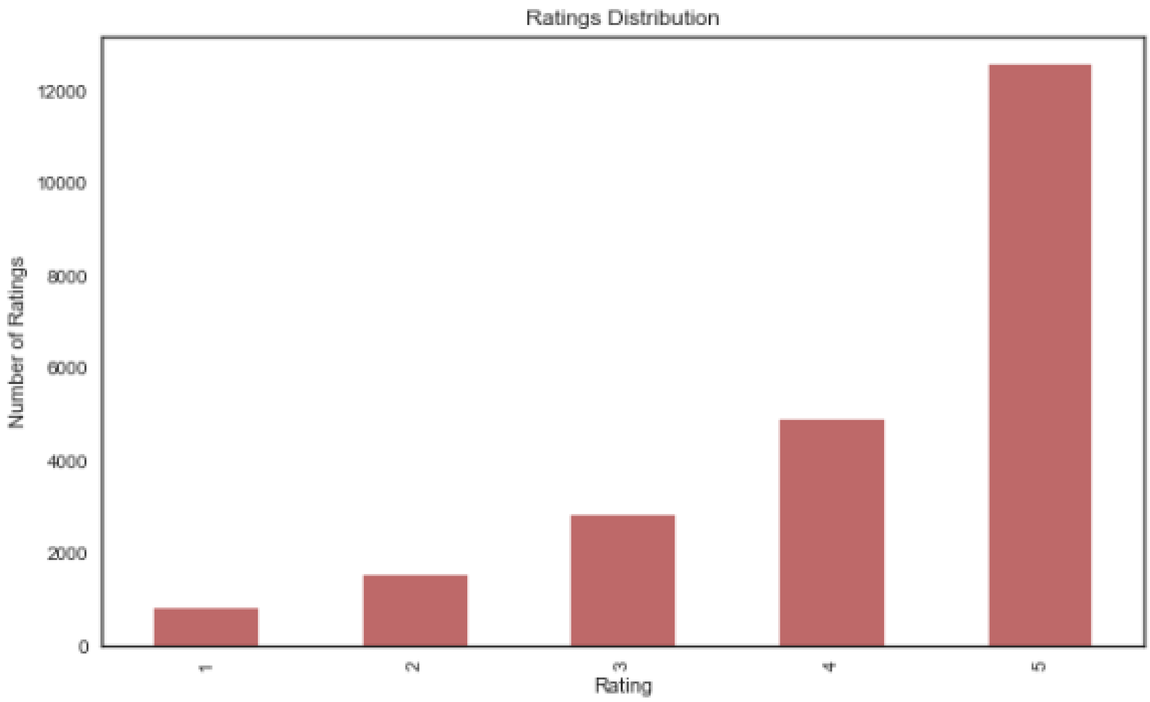
The distribution of reviews for different product types can be shown as follows.



Tops were the most reviewed product type followed by dresses and then bottoms. Trend categories of products are either least bought or least reviewed amongst the lot.

4.3 Rating

The distribution of ratings for all the product types show that most products were rated with a rating of 5.



5. Natural Language Processing on Reviews

After conducting basic Exploratory Data Analysis (EDA) on the dataset, the next phase of analysis delves deeper into understanding sentiment extraction and its association with other attributes. This involves Natural Language Processing (NLP) techniques to uncover underlying trends and insights within the customer reviews dataset.

5.1 Sentiment Extraction and Trends

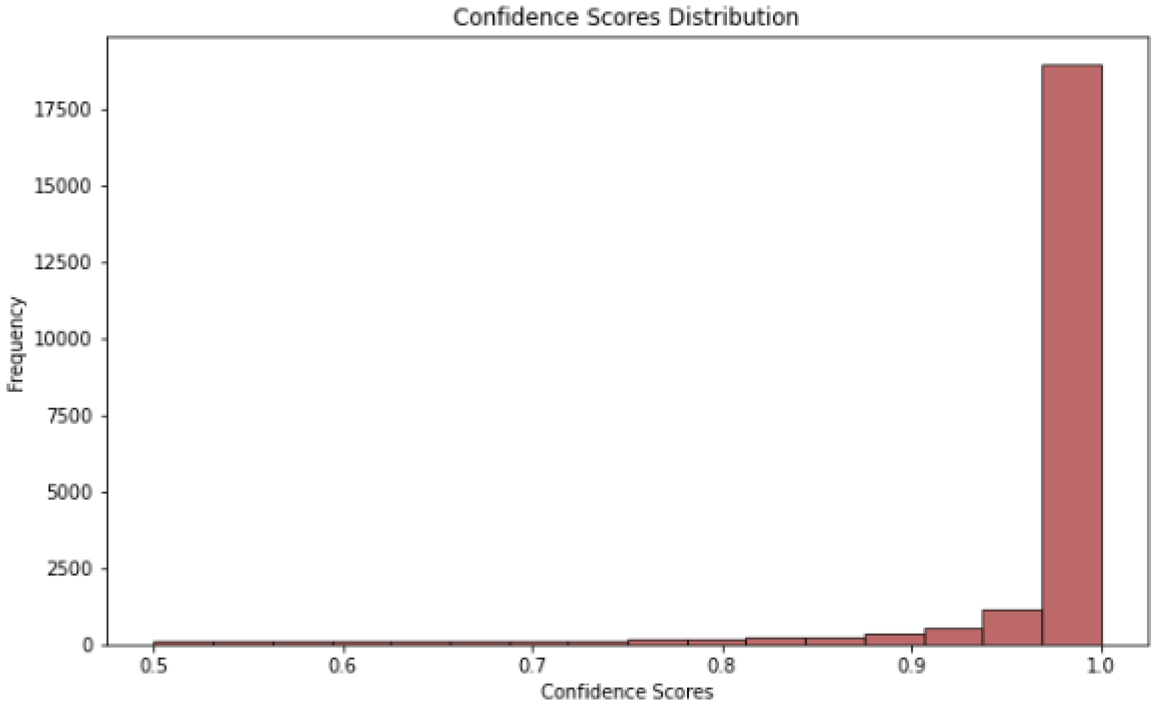
Utilizing sentiment analysis techniques, the aim is to extract the sentiment polarity (positive, or negative) from each customer review. This then allows to gauge overall customer satisfaction and discern any prevalent sentiment trends over time or across different product categories.

A BERT-based pre-trained model was used to extract the sentiments of each review within the dataset. Based on the requirement, DistilBERT was chosen because:

- It is a faster, smaller version of BERT, retaining 97% of BERT's performance while being more computationally efficient.
- It is optimized for sentiment analysis with high accuracy on English text.

- It handles text case-insensitively, improving robustness against varied text formatting in reviews.
- Ensures consistency and reproducibility of results.
- Additionally, the implementation process is relatively straightforward because the dataset is not annotated, thus it cannot be used for the task of fine-tuning the BERT model specifically for a dataset on reviews from an online retailer.

The model's confidence scores associated with its sentiment predictions per review provide valuable insights into the quality of sentiment predictions. Analyzing these confidence scores allows for a better understanding of the reliability and accuracy of the sentiment predictions made by the model. It is evident that the model exhibits a high level of confidence in its sentiment predictions, indicating a strong degree of certainty in its assessments. This is shown in the following distribution.



Sentiment Over Rating:
Analyzing the sentiment values over the ratings given for the same reviews show an interesting result.

Sentiment	NEGATIVE	POSITIVE
Rating		
1	709	112
2	1242	307
3	1835	988
4	1322	3586
5	857	11682

The ratings, often considered the benchmark for product comparison, do not perfectly represent the user's opinion. This is especially noticeable in products with ratings of 3 and 4. Ratings of 1 and 2 clearly show a dominance of negative sentiments, while ratings of 5 predominantly reflect positive sentiments.

Sentiment Over Age Groups:
Analyzing the sentiment values over different age groups give the following insights:

Sentiment / Age_Group	NEGATIVE	POSITIVE
0-19	7	37
20-29	699	2096
30-39	2059	5286
40-49	1595	4308
50-59	960	2874
60-69	533	1723
70-79	95	260
80-100	17	91

Across all age groups, the number of positive sentiments is higher than negative sentiments. This suggests that the majority of people have a positive outlook.

The age group '30-39' has the highest number of both negative and positive sentiments. This could indicate that this age group is more vocal or active in expressing their sentiments.

There is a decline in the number of sentiments (both positive and negative) as the age group increases from '50-59' onwards. This might suggest less participation from older age groups; however, since the number of reviewers in these age groups is generally lower, this cannot be definitively stated. The youngest ('0-19') and oldest ('80-100') age groups have the lowest sentiment counts. But this again likely results from less representation of these age groups in the data source.

Sentiment Over Product Types:
Analyzing the sentiment values over different different product types give the following insights:

Sentiment / Type	NEGATIVE	POSITIVE
Bottoms	946	2715
Dresses	1718	4427
Intimate	387	1266
Jackets	252	750
Tops	2612	7436
Trend	50	68

Comparing sentiment distributions across different product categories provides insights into customer preferences and satisfaction levels for each product type. Interestingly, the product type of tops have the most positive as well as negative reviews. This is followed by the product type of dresses.

Review Length:
Review length can play an important role in the analysis of customer feedback. Longer reviews offer richer insights into customer experiences and can differentiate between superficial and in-depth feedback. They may impact sentiment analysis, help detect spam, and indicate customer engagement. Additionally, there may be a correlation between review length and ratings, as highly satisfied or dissatisfied customers often write longer reviews.

Upon analysis of the length of reviews within this data:

- No significant trends observed between review length vs. ratings
- No significant trends observed between review length vs. product type

Correlation analysis results:

- Pearson correlation indicates a very weak negative linear relationship between review length and rating
- Spearman correlation indicates a very weak negative monotonic relationship between review length and rating

Generally, longer reviews might be associated with slightly lower ratings, but the relationship is not strong enough to draw definitive conclusions. While review length can provide some insights into customer feedback, it does not strongly correlate with the ratings given by customers.

5.2 Named Entity Recognition (NER)

In addition to sentiment analysis, Named Entity Recognition (NER) is employed to identify specific keywords or entities associated with negative reviews. By extracting relevant entities from negative reviews, common pain points and problem areas experienced by customers can be identified. These may include issues such as delivery problems (e.g., late delivery), product defects (e.g., damaged product), size discrepancies, or quality concerns.

The most effective approach for this task is to fine-tune a Named Entity Recognition (NER) model on a dataset that closely resembles the target dataset—in this case, an online retailer reviews dataset. By utilizing a dataset with annotated entities, the NER model can learn to extract specific issues mentioned in the reviews. This fine-tuning process ensures that the NER model is optimized for accurately identifying relevant entities within the context of online retailer reviews. This approach enhances the precision and relevance of the extracted issues, leading to more insightful analysis and actionable insights from the review dataset.

Given the extensive resources and annotated dataset or pre-trained model required for fine-tuning an NER model, simpler approaches can be adopted to address this task. These approaches offer effective alternatives for extracting insights from the review dataset without the need for complex modeling.

- **Keyword Matching Approach**:
  - This method involves creating a list of keywords or phrases related to the specific issues, such as late delivery, damaged items, size issues, etc.
  - The review texts are then scanned for the presence of these keywords using basic keyword matching techniques.
  - If any of the keywords are found in the review text, it indicates that the review mentions a particular issue.
  - This approach is relatively straightforward and does not require sophisticated NLP tools but may overlook nuanced mentions of issues that do not exactly match the predefined keywords.

- **spaCy's Named Entity Recognition (NER)**:
  - spaCy's NER functionality is a more advanced approach that automatically identifies named entities, such as dates, quantities, organizations, locations, and more, within the text.
  - Negative review texts are processed using spaCy's NER pipeline, which identifies entities relevant to potential reasons for negative feedback, such as late delivery, size issues, damaged products, etc.
  - Identified entities are then analyzed for frequency and distribution to identify commonly cited issues.
  - While spaCy's NER provides more comprehensive entity recognition, it may miss certain entities due to limitations in the trained model.

- **Topic Modeling using LDA**:
  - Topic modeling, particularly using algorithms like Latent Dirichlet Allocation (LDA), is another method for identifying common themes or topics within a collection of documents, including negative reviews.
  - Negative review texts are preprocessed to remove noise and irrelevant information, then fed into the LDA algorithm to identify latent topics.
  - Each identified topic consists of a set of keywords that represent common issues or concerns expressed by customers in negative reviews.
  - Topic modeling provides a broader understanding of the prevalent themes across negative reviews but may require more computational resources and expertise in NLP.

This analysis is presented in the next section of Analysis under the subsection of Review Analysis.

6. Analysis

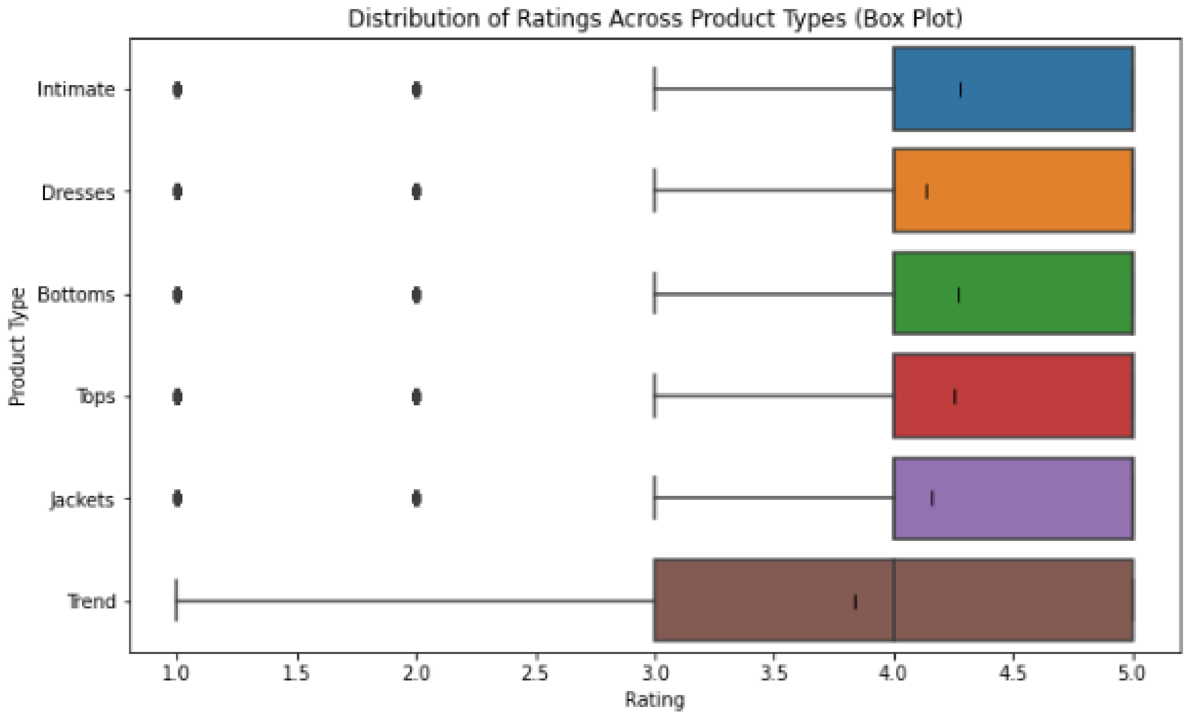
The following section details on the analysis focused on answering the earlier defined problem statements. Each of the following subsections talk about analysis focused on different attributes or groups of attributes within the dataset.

Product Type and Ratings

Based on the type of the product id, its type and the ratings associated with them, the following questions can be asked:

- **Which products are most reviewed?**
- **Which products are most popular under each type?**
- **What are the most frequently reviewed product types?**
- **How do their ratings compare?**
- **What is the distribution of ratings across all product types?**

The product type of tops are most frequently reviewed. This can be observed in the product type distribution plot as shown in section 4.2. The average rating across each of the product types is shown in the following distribution of the ratings across the different product types.

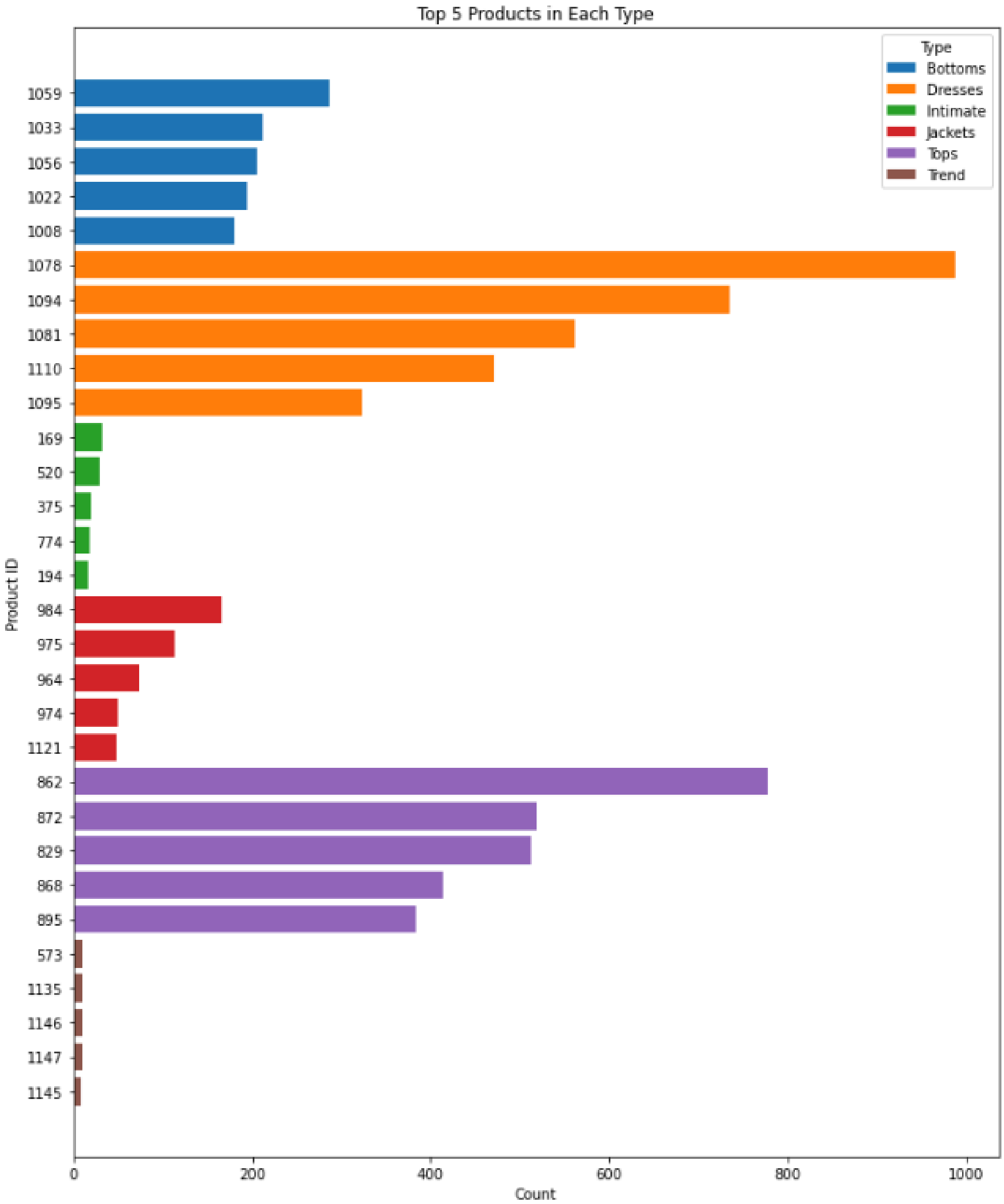


The top 10 product_ids with the most reviews are as follows:

	Product_ID	Count
0	1078	987
1	862	778
2	1094	735
3	1081	561
4	872	519
5	829	512
6	1110	471
7	868	414
8	895	384
9	936	348

Product_ID 1078 has the most reviews out of all.

The 5 most reviewed product_ids within each type are as follows. Out of this, it can be noted that Product_ID 1078 does have the most number of reviews, which is naturally the most reviewed under the category of dresses. The product ids can be used to identify the associated reviews to identify what went well which led to their popularity.

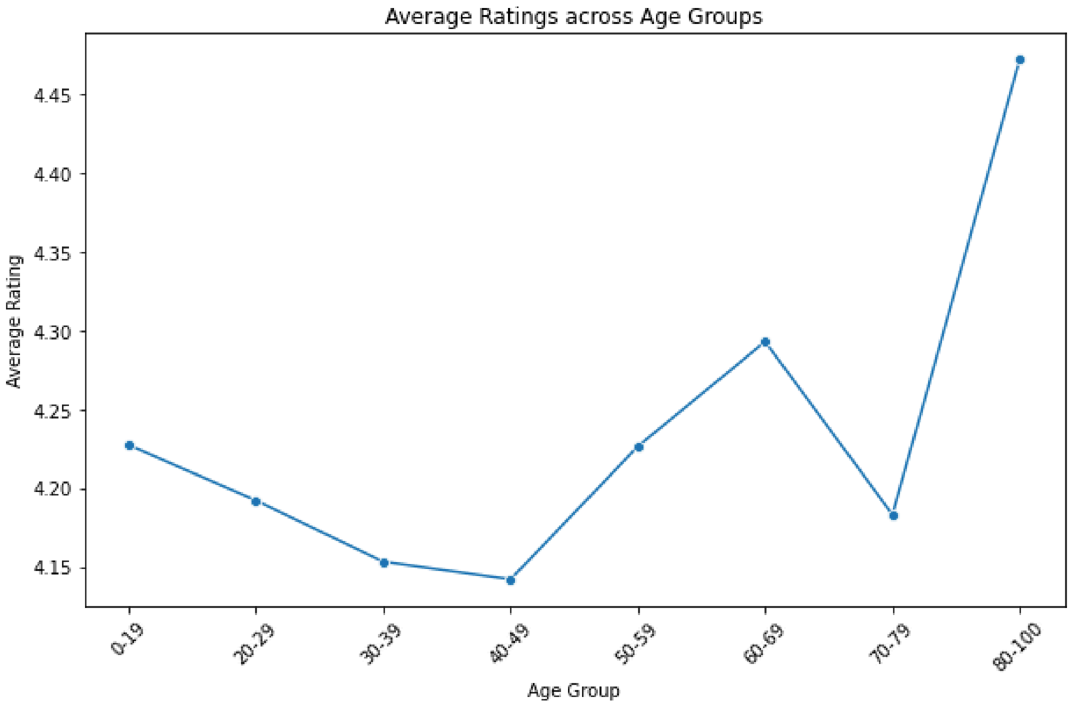


Age Groups, Rating and Reviews

Based on different age groups, the ratings associated with them as well as the reviews, the following questions can be asked:

- **Are there any noticeable trends in ratings and reviews based on age groups within each product type?**
- **Do certain product types receive consistent ratings across different age groups?**
- **How does age relate to changes in ratings and types of products reviewed? Does age of a person affect the rating?**
- **How does review sentiment relate to the ratings given?**
- **Are there any noticeable trends in sentiment of reviews based on age groups within each product type?**
- **What is the sentiment for each type of product? Can it be quantified with a rating? Are there any discernible patterns?**
- **What are the top reviewed products within different age groups?**

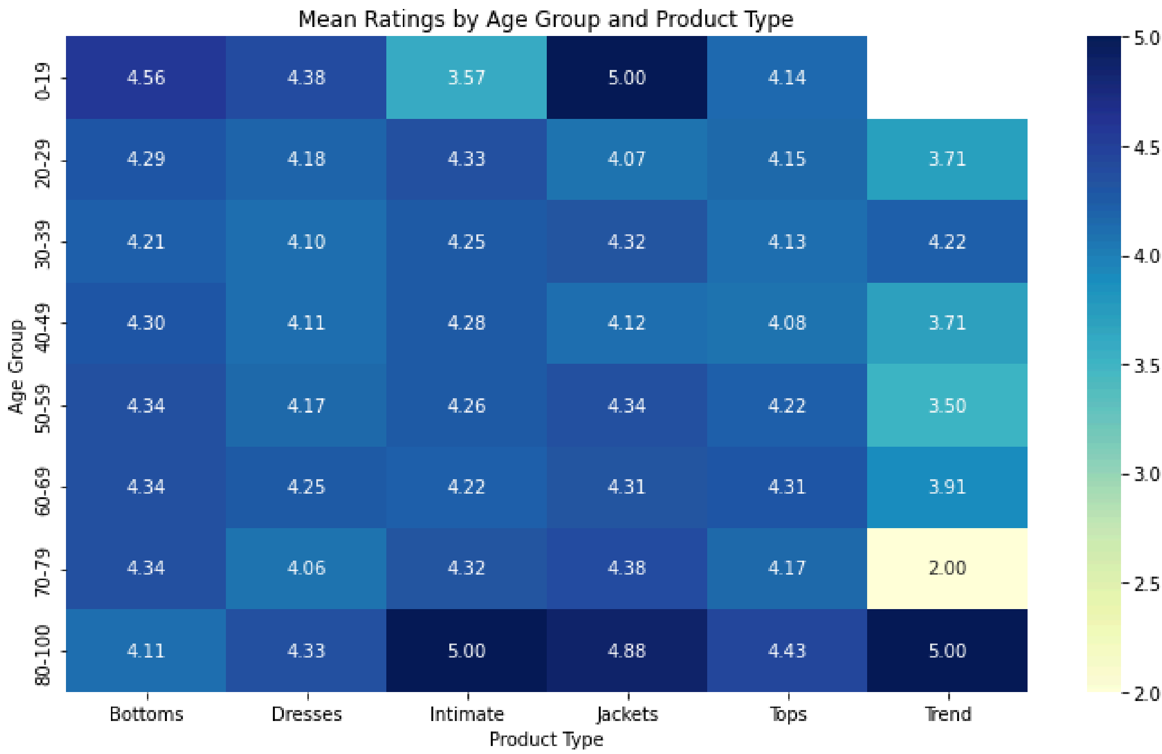
Analyzing average ratings across all the age groups gives the following trend plot.



The average rating provided by the age group of 80-100 is the highest, while the age group 40-49 yields the lowest average rating. This suggests that satisfaction levels vary across age groups, with the oldest age group appearing the most satisfied. However, it's essential to note that **the relationship between age and ratings is not necessarily linear across all age groups**. Additionally, the relatively low representation of the age group 80-100 in the dataset could influence these findings, potentially skewing the observed satisfaction levels. **Since most reviewers are in the age groups of 30-49, the average rating within them is similar. This again is heavily influenced due to the bias caused by the low-representation of certain other age groups.** Accurate analysis can be performed if the data is not biased or if the extreme age groups are eliminated from the analysis. Then there will be an increase in the mean rating with age until the age group of 60-69.

Mean ratings across age groups as well as product types can be visualized using a heatmap. The following observations can be made:

- **Products under the category of tops, bottoms as well as dresses have a consistent rating across all the age groups.**
- **Products under the Trend category have a varied mean rating across all the age groups.**
- The age group of 80-100 have a consistently high average rating across all the product types with Intimate products and Trend products having a mean rating of 5. The issue of low representation of this age group could have had an effect here.
- The product type with the worst rating is that of Trend within the 70-79 age group.
- The age group of 0-19 did not review any products of the type trend and that of intimate were rated low. This age group preferred jackets the best.

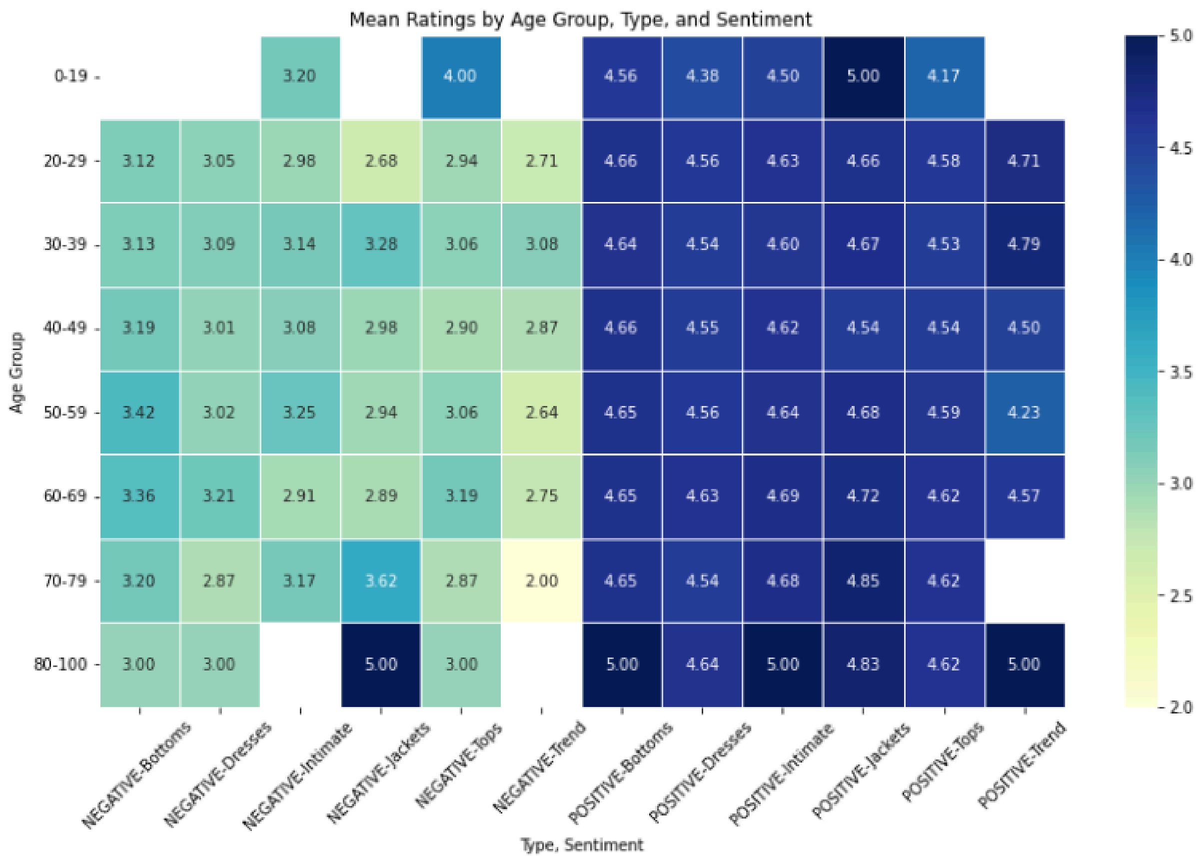


Analyzing reviews across age groups and product types and mean ratings. Review sentiments are obtained through sentiment analysis of the text reviews. The following observations can be made as per the heatmap:

- **The average rating to the right is higher (blue) which corroborates the positive sentiment of the reviews.** This is valid across all product types as well as age groups.
- **The average rating to the left is lower (green) which corroborates the negative sentiment of the reviews.** This is valid across all but two product types-age groups category.
- The mean rating for tops under the 0-19 age group is high in the range of 4 but the sentiment states Negative, which is a contradiction.
- The mean rating for jackets within the 80-100 age group is notably high, standing at a value of 5. However, the sentiment analysis indicates a negative sentiment, which presents a contradiction. This discrepancy could be attributed to various factors, such

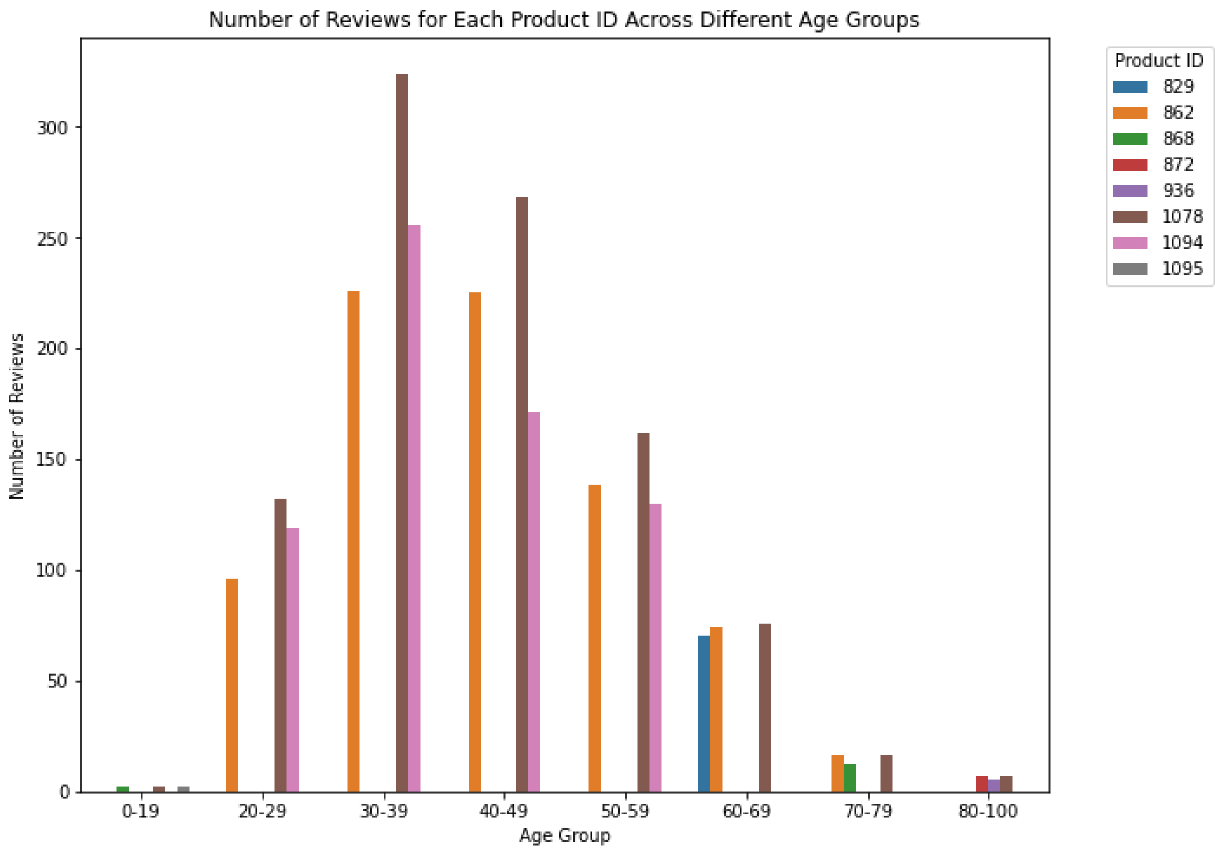
as misinterpretation of the review content or incorrect recording of the rating. It underscores the importance of thorough data validation and review when analyzing sentiment and ratings to ensure accurate insights.

- The same is also observed with the age group of <19, wherein Tops are rated high despite the review being negative in nature. Which is another example of misinterpretation of the review content or incorrect recording of the rating. Point to note here is that this is again an underrepresented group with less than 0.2% of total reviewers.



The conclusions drawn from this study of associating reviews with age groups and product types are consistent with those drawn from associating ratings with age groups and product types. The issue of misrepresentation due to lesser data pertaining to certain age groups can cause ambiguities in the trends. But apart from that, **product type of Trend does have lower rating and review in general across all age groups and products under the category of tops, bottoms as well as dresses have a consistent rating across all the age groups.**

From the following plot of number of reviews for top three most reviewed products across different age groups, it is evident that **product id 1078 is the most reviewed across all the age groups consistently**. Product 1094 and Product 862 come close in 4 of the age groups.



Product Type, Rating and Reviews

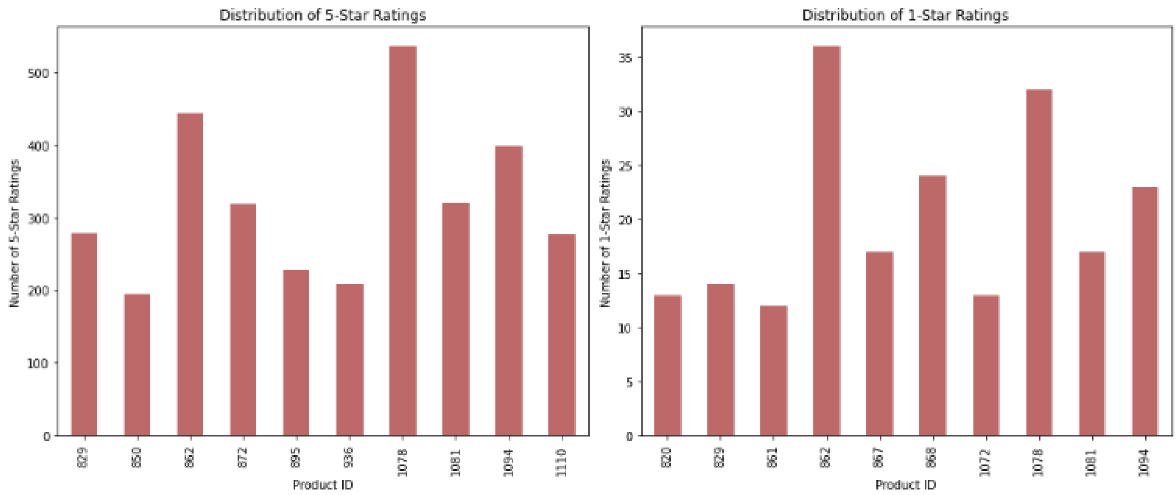
Based on different product type, the ratings associated with them as well as the reviews, the following questions can be asked:

- **Is there a particular product or product type that consistently performs the best or the worst in terms of ratings and reviews?**
- **What are the most frequent review words for each product type?**

Identifying the most 5 rated and most 1 rated product and product type can help determine if it performs well or worse.

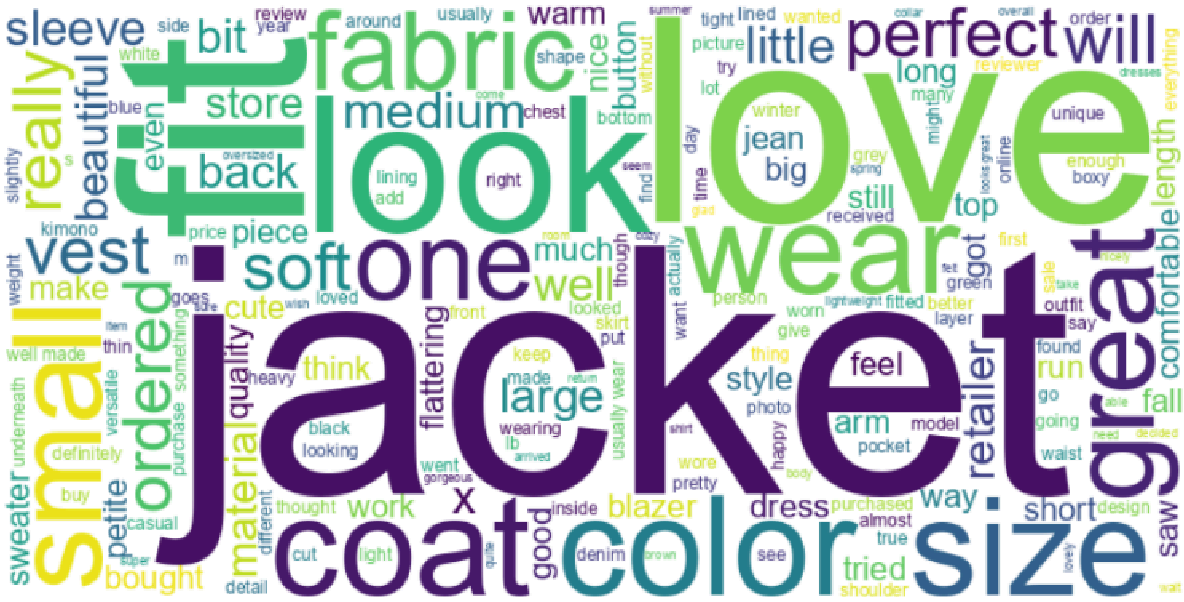
Upon analyzing ratings and reviews, **Product ID 1078 consistently garners 5 ratings, while Product ID 862 consistently lags behind with the most 1 ratings**. From the top 5 products under each type plot, it can be noted that Product ID 1078 is categorized under dresses whereas Product ID 862 is categorized as Top.

Similarly, specific product types stand out. Tops consistently earn the highest number of 5 ratings, reflecting exceptional customer satisfaction. Ironically, Tops consistently receives the lowest number of 1 ratings as well. This could be because as per the number of items bought within each product type, Tops are the most bought which is significantly larger than the others.



The most frequent words within the reviews for each product type can be found using NLP techniques including tokenizing, removing stop words and then counting its frequency. This can also be visualized by using WordCloud.

A sample WordCloud for product type of Jackets:



The top 10 keywords within each product type are shown in the following figure:

Intimate

	Word	Frequency
0	love	605
1	size	589
2	like	490
3	wear	487
4	fit	463
5	great	436
6	small	415
7	would	385
8	top	370
9	soft	347

Dress

	Word	Frequency
0	dress	9432
1	size	2606
2	love	2312
3	fit	2101
4	like	1890
5	wear	1768
6	5	1754
7	would	1681
8	fabric	1631
9	great	1450

Bottom

	Word	Frequency
0	fit	1793
1	size	1767
2	love	1551
3	skirt	1257
4	pants	1247
5	great	1177
6	like	1149
7	5	1057
8	wear	1056
9	jeans	997

Top

	Word	Frequency
0	top	5710
1	love	3976
2	size	3315
3	like	3071
4	wear	2758
5	great	2722
6	fit	2588
7	shirt	2320
8	would	2315
9	color	2250

Jacket

	Word	Frequency
0	jacket	629
1	love	449
2	size	365
3	like	356
4	wear	314
5	coat	298
6	great	286
7	small	285
8	fit	282
9	would	243

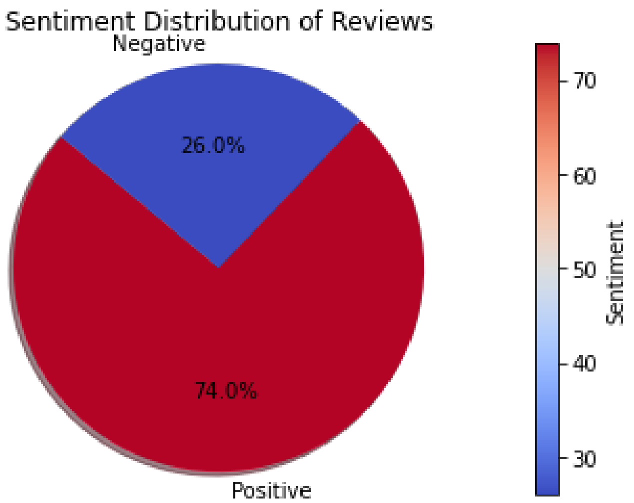
Trend

	Word	Frequency
0	dress	86
1	size	55
2	like	52
3	fit	44
4	fabric	39
5	love	38
6	would	37
7	look	35
8	wear	33
9	beautiful	32

Reviews Sentiment

Based on the reviews and their derived sentiments, the following questions can be asked:

- **What is the overall sentiment of text reviews?**
- **Does the age of the customer influence the words used in the reviews?**



The sentiment analysis of the text reviews reveals a predominantly positive sentiment, with approximately 74% of the reviews reflecting positive sentiment. This suggests that the majority of the reviews express favorable opinions or experiences.

This can be analyzed by identifying the most frequently used words within the reviews as per different age groups. This can also be visualized using WordClouds. The most frequently used keywords is presented as a table following the WordClouds.

A sample WordCloud for the age group of 20-29:

A sample WordCloud for the age group of 60-69:

Age Group : 0-19

	Word	Frequency
0	dress	29
1	like	20
2	top	17
3	size	16
4	really	15
5	small	15
6	well	13
7	little	13
8	fit	13
9	wear	13

Age Group : 20-29

	Word	Frequency
0	dress	1592
1	love	1131
2	size	1016
3	top	943
4	fit	897
5	like	825
6	wear	811
7	would	732
8	great	731
9	5	611

Age Group : 30-39

	Word	Frequency
0	dress	3508
1	love	2873
2	size	2753
3	like	2443
4	fit	2419
5	top	2270
6	great	2036
7	wear	2000
8	would	1877
9	5	1667

Age Group : 40-49

	Word	Frequency
0	dress	2660
1	love	2345
2	size	2309
3	top	1868
4	fit	1837
5	like	1782
6	wear	1678
7	great	1596
8	would	1348
9	5	1313

Age Group : 50-59

	Word	Frequency
0	dress	1677
1	size	1495
2	love	1452
3	top	1295
4	fit	1215
5	like	1156
6	wear	1092
7	great	992
8	5	914
9	fabric	877

Age Group : 60-69

	Word	Frequency
0	love	922
1	size	902
2	dress	839
3	top	800
4	fit	770
5	wear	686
6	like	667
7	great	618
8	fabric	586
9	color	489

Age Group : 70-79

	Word	Frequency
0	size	143
1	dress	142
2	love	140
3	wear	122
4	top	112
5	fit	95
6	color	94
7	great	94
8	small	86
9	ordered	84

Age Group : 80-100

	Word	Frequency
0	size	65
1	love	62
2	dress	50
3	ordered	46
4	top	43
5	like	36
6	beautiful	31
7	color	30
8	would	30
9	soft	30

The WordClouds as well as the frequently used keywords tables do not show any trends of certain words being used by certain age groups. Irrespective of the age groups, one of the most frequent keywords is dress. A point to be noted is that with increasing age, of more than 60, the emphasis of the reviews move towards size and expression of likeness toward the product.

Review Analysis

Based on the analysis of the reviews, the following questions can be asked:

- **Are there specific issues commonly mentioned in negative reviews as:**
    a. **Delivery problems (e.g., late delivery)**
    b. **Damaged product**
    c. **Size discrepancies**
    d. **Color not matching the product description**

e. **Size inclusivity**
f. **Quality of the product (stitching issues/ torn product)**

- **Are there specific keywords commonly mentioned in the positive reviews?**

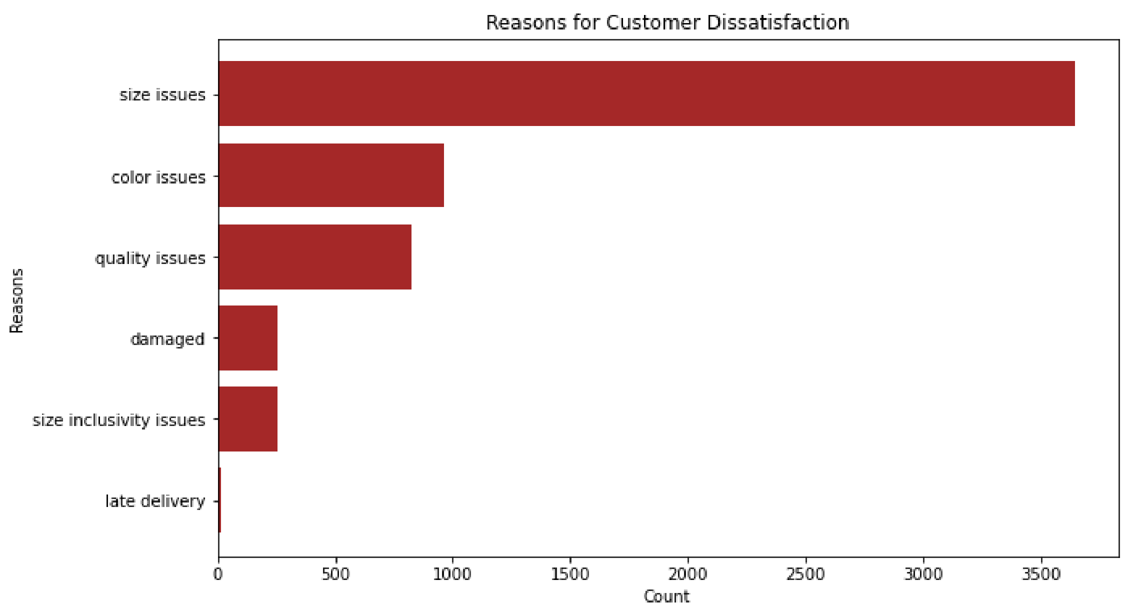
To address this question, of identifying **issues commonly mentioned in negative reviews**, several methodologies could be employed. One viable approach involves scrutinizing negative sentiment reviews to discern recurring keywords. These keywords can then be categorized into specific issues such as late delivery, damaged goods, size discrepancies, color mismatches, quality concerns, and inclusivity issues. By systematically categorizing negative feedback, the retailer can pinpoint areas for improvement, facilitating targeted enhancements to the customer experience.

Keyword Matching
Implementing the basic keyword matching technique to search for relevant terms or phrases in the text of negative reviews. This involves creating a list of keywords related to each issue (e.g., "late delivery," "damaged," "size issues") and checking if any of these keywords appear in the review text.

The following keywords were associated with the identified potential reasons for negative reviews:

- Late Delivery: Late, Delay, Delayed, Delivery
- Damaged Items: Damaged, Torn, Ruined, Delivery, Stitching, Stitches, Holes
- Size Issues: Size, Fit, Fitting, Small, Large, Tight, Loose
- Color Issues: Color, Colour, Hue, Shade
- Quality Issues: Quality, Poor, Bad
- Size Inclusivity Issues: Inclusive, Options, Sizes

Doing this type of assessment, led to observing that **size issues** were the most commonly mentioned reason within the negative reviews. This can be observed by the frequency plot obtained from the keyword mapping analysis.

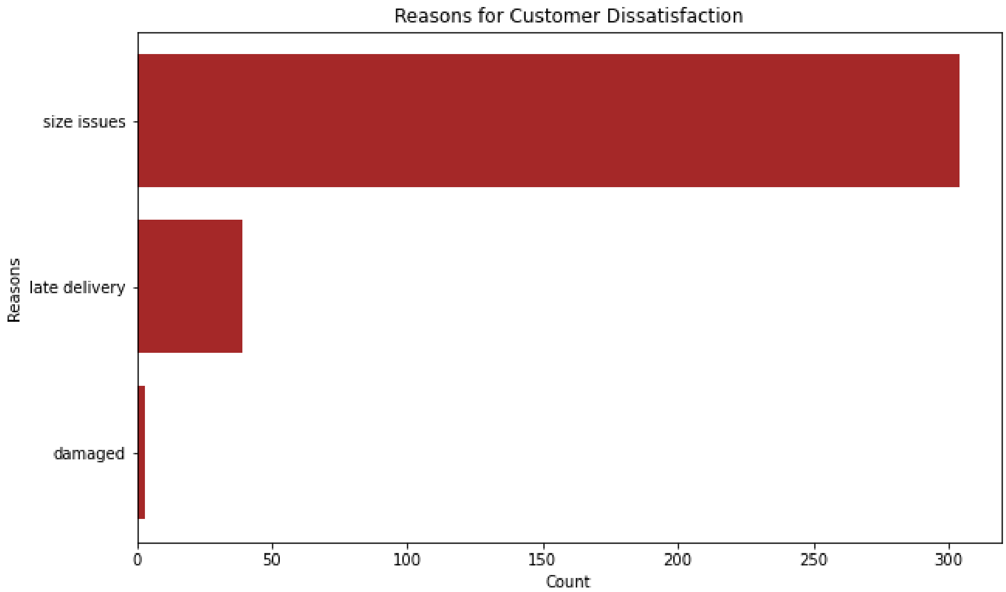


spaCy's Named Entity Recognition

The second type of analysis was done using spaCy's Named Entity Recognition (NER). This was utilized to extract relevant entities from negative reviews. Leveraging spaCy's NER functionality, entities within the text were identified. Identified entities include dates, quantities, and other entities to determine potential reasons for negative feedback. The analysis then delved into the frequency and distribution of these entities, to identify the commonly cited issues. The issues searched for in this method include:

- Late delivery
- Size issues
- Damaged products
- Color issues
- Quality issues

Out of these, the algorithm could only identify three, which is mainly due to the shortage of the recognized entities within spaCy's NER model. Yet, the reason highlighted here was that of **size issues**.



Topic Modeling using LDA

The third type of analysis was done based on topic modeling. Topic modeling is a technique used in natural language processing (NLP) to discover latent topics or themes within a collection of documents. One popular algorithm for topic modeling is Latent Dirichlet Allocation (LDA). The primary goal of topic modeling is to automatically identify and extract underlying patterns of topics from text data without prior labeling or annotation.

In this analysis, LDA is used to uncover hidden themes or topics present in the corpus of negative reviews. This can help to gain valuable insights into the common issues or concerns expressed by customers across various reviews.

The process of conducting topic modeling with LDA involves several steps. Firstly, the text data, in this case, negative reviews, is preprocessed to remove noise and irrelevant information, such as stop words and punctuation. Next, the LDA algorithm is applied to the preprocessed text data to identify topics and their associated keywords. Finally, the results of the topic modeling are interpreted and analyzed to understand the prevalent themes or issues within the negative reviews.

This analysis identified the following topics:

- Topic 1: quality, fabric, look, like, design, dress, price, cute, poor, material
- Topic 2: size, dress, color, small, fit, like, ordered, look, wear, medium
- Topic 3: size, dress, fit, small, like, shirt, wear, look, love, ordered
- Topic 4: blocking, boring, disproportionate, paper, harder, robe, frankly, select, chemise, sooooo
- Topic 5: dress, look, like, fabric, fit, love, material, color, sweater, really

The following themes can be observed using the obtained results:

- Topic 1: This topic revolves around aspects related to the quality and design of products, particularly clothing items like dresses. Keywords such as "quality," "fabric," "look," and "design" indicate discussions about the overall appearance and feel of the garments. Terms like "poor" and "material" suggest that customers may be expressing dissatisfaction with certain aspects of product quality.
- Topic 2: This topic is centered on issues related to the sizing and fit of dresses. Keywords like "size," "dress," "color," and "fit" suggest discussions about how well the garments match customers' expectations in terms of size and fit. Terms like "small," "ordered," and "medium" indicate specific concerns about sizing options and the accuracy of product descriptions.
- Topic 3: Similar to Topic 2, this topic also focuses on sizing and fit issues, particularly with shirts and dresses. Keywords such as "size," "dress," "fit," and "small" indicate ongoing discussions about the appropriateness of clothing sizes and how they align with customers' preferences. Terms like "love" and "ordered" suggest that despite sizing concerns, customers may still have positive sentiments about the products.
- Topic 4: This topic appears to encompass discussions about miscellaneous issues that customers have encountered with their purchases. Keywords like "blocking," "boring," and "disproportionate" suggest dissatisfaction with certain product attributes or features. Terms like "paper," "harder," and "frankly" could indicate frustrations or criticisms regarding the perceived quality or utility of the items.
- Topic 5: This topic captures general sentiments and preferences expressed by customers regarding dresses and other clothing items. Keywords like "dress," "look," "fabric," and "fit" suggest discussions about the overall appeal and suitability of the products. Terms like "love," "material," and "color" indicate positive sentiments and preferences for certain styles or designs.

Based on the analysis conducted, it is conclusive that specific issues commonly mentioned in negative reviews include size discrepancies and fit problems, dissatisfaction with the quality and design aspects of the products, and various miscellaneous issues such as perceived product flaws and boredom with certain styles. Among these, **size issues emerge as the most prevalent concern among customers, with a significant number of negative reviews expressing dissatisfaction with the sizing accuracy and consistency across product listings**. Additionally, customers often highlight issues related to fabric quality, overall appearance, and perceived value for the price paid, particularly in clothing items like dresses.

In conclusion, the observed topics underscore the importance of closely monitoring customer feedback and addressing areas of concern to enhance product quality, sizing accuracy, and overall customer satisfaction. By leveraging insights from topic modeling analysis, retailers can identify actionable opportunities for product improvement, thereby fostering better relationships with customers and driving long-term success in the marketplace.

The most common keywords within the positive reviews can be visualized by a WordCloud as follows:

7. Conclusion

Currently, as per the negative reviews, the main issues include:

1. **Size discrepancies and fit problems:**
   Customers express dissatisfaction with the sizing accuracy and consistency across product listings within the negative reviews. This indicates that products were received that do not fit as expected, leading to inconvenience and potential returns or exchanges.
2. **Dissatisfaction with quality and color:**
   Negative reviews also often mention concerns regarding fabric quality and color. Customers received products that did not meet their expectations in terms of color or quality or fabric, and it can lead to disappointment and dissatisfaction.

Addressing these points is crucial for the retailer to enhance product quality, sizing accuracy, and overall customer satisfaction, thereby fostering better relationships with customers and driving long-term success in the marketplace.

Based on the positive reviews, it's evident that customers are delighted with various aspects of the products. They express love for the items, highlighting their overall satisfaction. The products are described as great, indicating excellence in quality, design, or fit. Customers are particularly pleased with the fit and size accuracy, suggesting that the products align well with their expectations. Tops stand out positively, possibly indicating exceptional quality or design in this category. Customers find the products enjoyable to wear, emphasizing comfort and style. They express a favorable opinion, indicating a general likeness towards the products. The color options are appreciated, suggesting satisfaction with the available choices and the vibrancy or accuracy of the colors. Overall, customers consistently rate the products highly, reflecting their overall satisfaction and positive experiences.

Considering both the pain points and the positive feedback, it's evident that improvements in several areas could enhance the overall customer experience.

- **Ensuring better representation of products on the retailer's website is crucial.** Clear descriptions regarding color options and sizing, accompanied by high-quality images, can help customers make informed decisions and set accurate expectations regarding fabric and quality. Implementing basic machine learning algorithms could further assist customers by suggesting the most suitable fit based on data from previous customers with similar body types.
- **Moreover, prioritizing the quality of products is essential to address concerns raised in negative reviews regarding fit, color, and overall quality.** This quality assurance process should start from the sourcing of raw materials, ensuring high standards are maintained throughout production. Providing transparency regarding the origin of raw materials can also foster trust among customers, reinforcing their confidence in the brand.

Appendix

This section contains all the other analyses that were tried including models and methods.

To obtain bins for different distributions, different methods were applied, but since the attribute in question is “Age”, the values given by these methods did not seem to do well. Hence 20 was chosen, which is close to the value obtained from Sturges’ Rule, 16.

In [85]:
```
# Sturges' Rule
no_bins = int(math.ceil(math.log2(len(data['Age'])) + 1))
no_bins
```
Out[85]: 16

In [86]:
```
# Square root rule
no_bins = int(math.sqrt(len(data['Age'])))
no_bins
```
Out[86]: 153

In [87]:
```
# Freedman-Diaconis Rule
q25, q75 = np.percentile(data['Age'], [25, 75])
bin_width = 2 * (q75 - q25) * len(data['Age'])**(-1/3)
no_bins = int((data['Age'].max() - data['Age'].min()) / bin_width)
no_bins
```
Out[87]: 64

In [88]:
```
# Rice Rule
no_bins = int(2 * len(data['Age'])**(1/3))
no_bins
```
Out[88]: 57

The NLP library FLAIR was used to evaluate its effectiveness along with that of Word2Vec, BERT and then DistilBERT to identify sentiment of the reviews.

In [90]:

```python
from flair.models import TextClassifier
from flair.data import Sentence

classifier = TextClassifier.load('en-sentiment')

sentence = "I bought a pair of socks and a blue skirt. It fit like a glo

sentence = Sentence(sentence)

classifier.predict(sentence)

print(sentence.labels[0])
```

```
Sentence[16]: "I bought a pair of socks and a blue skirt. It fit like a glo
ve" → POSITIVE (0.9822)
```

Since Word2Vec does better with words and BERT is the current state of the art for grasping the context within a group of text, BERT was focused on. Due to the aforementioned advantages to using DistilBERT within this context, it was used.

Identifying reasons for negative reviews initially involved attempting NER methods. Despite the availability of a pre-trained NER model tailored for the fashion and luxury industry [https://huggingface.co/AkimfromParis/NER-Luxury] , access to it necessitated approval from its authors, which was not obtained in a timely manner.

Subsequently, the focus shifted towards fine-tuning a BERT model using existing fashion/clothing/online retailer review datasets. However, this process proved to be time-consuming due to the extensive size of the Amazon Fashion Dataset utilized for fine-tuning.

In [14]:

```python
import gzip
import json
import pandas as pd

file_path = 'Amazon_Fashion.jsonl.gz'

data = []

with gzip.open(file_path, 'rt', encoding='utf-8') as file:
    for line in file:
        data.append(json.loads(line))
df = pd.DataFrame(data)
```

```
   rating                title  \
0     5.0         Pretty locket
1     5.0                     A
2     2.0             Two Stars
3     1.0        Won't buy again
4     5.0  I LOVE these glasses

                                                text images        asin  \
0  I think this locket is really pretty. The insi...     []  B00LOPVX74
1                                              Great     []  B07B4JXK8D
2  One of the stones fell out within the first 2 ...     []  B007ZSEQ4Q
3  Crappy socks. Money wasted. Bought to wear wit...     []  B07F2BTFS9
4  I LOVE these glasses!  They fit perfectly over...     []  B00PKRFU4O

  parent_asin                       user_id      timestamp  helpful_vote  \
0  B00LOPVX74  AGBFYI2DDIKXC5Y4FARTYDTQBMFQ  1578528394489             3
1  B07B4JXK8D  AFQLNQNQYFWQZPJQZS6V3NZU4QBQ  1608426246701             0
2  B007ZSEQ4Q  AHITBJSS7KYUBVZPX7M2WJCOIVKQ  1432344828000             3
3  B07F2BTFS9  AFVNEEPDEIH5SPUN5BWC6NKL3WNQ  1546289847095             2
4  B00XESJTDE  AHSPLDNW5OOUK2PLH7GXLACFBZNQ  1439476166000             0
```

Additionally, POS tagging in negative reviews was explored as another avenue to uncover reasons for negative reviews. This approach involved analyzing nouns, adjectives, and verbs within reviews to identify objects, describe issues' severity, and highlight problematic actions or events. Word2Vec was trained to further enhance this method, and SpaCy was also employed in the process. Once keywords were identified through POS tagging, their cosine similarity to the problems identified and their keywords can be used to classify the issues, offering a potential method for issue classification. However, despite its potential, this method was also found to be time-consuming and insufficient in addressing the problem statement adequately.

In [89]:
```
import nltk
from nltk import word_tokenize, pos_tag

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')

text = "I bought a pair of socks and a blue skirt. It fit like a glove"

tokens = word_tokenize(text)

tagged_tokens = pos_tag(tokens)

nouns = [word for word, pos in tagged_tokens if pos.startswith('NN')]

print(nouns)

```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\aishw\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\aishw\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
['pair', 'socks', 'skirt', 'glove']
```

Extracting Nouns looked like this within the reviews for the product type “Trend”:

In [137]:
```
for i in trend_df.index:
    print(extract_entites(str(trend_df['Title'].loc[i] )+ str(trend_df['Text_Review'].loc[i])))
#     break
```

```
[]
['dress', 'tie', 'buttons', 'feel']
[]
['tee', 'i', 'review', 'tee', 'jessa', 'tee', 'mail', 'today', 'i', 'deal', 'tee', 'opinion', 'reviewer', 'cotton', 'fabr
ic', 'fact', 'sweater', 'everything', 'tee', 'elbow', 'neckline', 'fit', 'everything', 'jeans', 'slacks']
['runs', 'size', 'fits', 'medium', 'bit', 'aline', 'knit', 'navy', 'i', 'elbow', 'length', 'sleeves', 'i', 'bust', 'fit',
'top', 'waist', 'basic', 'someone', 'shorter', 'length']
[]
['color', 'sale']
[]
['skirt', 'i', 'starters', 'waist', 'inches', 'front', 'bueno', 'gripe', 'color', 'nothing', 'picture', 'ivory', 'top',
'background', 'color', 'skirt']
['anything', 'closet', 'drape', 'fabric', 'color', 'placket', 'inside', 'sleeves', 'medium', 'length', 'drape', 'fit', 's
ize']
[]
[]
['dress', 'price', 'tts', 'me-', 'i', 'lbs.', 'body', 'issues', 'neck', 'braless', 'bra', 'material', 'light', 'breezy',
'dress', 'retailer', 'purchase', 'time']
['size', 'store', 'fun', 'petite', 'chance', 'order', 'size', 'i', 'reviews', 'dress', 'order', 'size', 'i', 'size', 'i',
```

In search of a more efficient solution, a BERT model fine-tuned specifically for reviews datasets was identified [https://huggingface.co/ongaunjie/distilbert-cloths-sentiment], although concerns regarding its reliability prompted the decision to proceed with DistilBERT instead.