Sentiment Analysis & Product Insights from Women's Clothing E-Commerce Reviews

SUMMARY

The project focused on analyzing customer sentiment for a Women's Clothing E-Commerce dataset containing over 23,000 reviews. By constructing a Natural Language Processing (NLP) pipeline, raw text data was cleaned, normalized, and scored for sentiment using both TextBlob and VADER algorithms. The analysis revealed that customer sentiment is overwhelmingly positive (over 93%), driven largely by the "Tops" and "Dresses" departments.
However, text analysis on negative reviews highlighted specific issues regarding sizing ("small", "big") and material quality ("fabric", "cheap"). Finally, the processed data was structured and serialized to generate a SQL script for database migration.

STEPS TAKEN FOR ANALYSIS

1. Data Ingestion and Cleaning

- **Loading:** Imported the raw dataset (Women's Clothing E-Commerce Reviews.csv) and inspected the schema (23,486 rows).
- **Null Handling:** Identified missing values and dropped rows where the Review Text was empty to ensure analysis integrity.
- **Text Cleaning:**
    - Converted all text to lowercase to standardize inputs.
    - Removed URLs and hyperlinks using Regular Expressions (Regex).
    - Removed HTML
    - Stripped non-alphanumeric characters (punctuation and numbers) and excessive whitespace.

2. Linguistic Processing (NLP Pipeline)

- **Tokenization:** Split raw review strings into individual words (tokens) using NLTK.
- **Stop Word Removal:** Filtered out common English words (e.g., "the", "is", "and") that provide little semantic value to sentiment analysis.
- **POS Tagging:** Applied Part-of-Speech tagging to understand the grammatical context of words (nouns, verbs, adjectives).

- **Normalization:** Applied **Lemmatization** (reducing words to their dictionary base form) to aggregate similar word forms.

3. Sentiment Classification

- **Model Application:** Applied two distinct sentiment analysis libraries to the cleaned text:
  - **TextBlob:** Calculated polarity scores.
  - **VADER (Valence Aware Dictionary and sentiment Reasoner):** Calculated compound sentiment scores.
- **Categorization:** Logic was applied to bucket scores into "Positive," "Neutral," or "Negative" labels for easy reporting.

4. Rating Analysis

- Converted rating to sentiment labels.
- Categorized rating as "Positive," "Neutral," or "Negative" labels for easy reporting.
- Compared rating sentiment and text sentiment.

5. Exploratory Data Analysis & Visualization

- **Frequency Analysis:** Visualized the most common words associated with each sentiment category using bar charts. This identified specific keywords driving negative feedback (e.g., "zipper", "fabric").
- **Category Performance:** Filtered the data for high ratings (4 and 5 stars) and grouped them by Department Name to identify the best-performing product categories (Tops and Dresses).

6. Data Engineering & SQL Preparation

- **Feature Extraction:** Calculated words count for every review.
- **Serialization:** Converted Python lists (tokenized data) into JSON strings to ensure compatibility with SQL storage formats.
- **SQL Generation:** Wrote a Python function to automatically generate a .sql file (Casestudy3database.sql) containing the CREATE TABLE schema and INSERT statements for the processed data.

INSIGHTS AND BUSINESS RECOMMENDATIONS

Percentage of Positive and Negative Reviews

Based on the sentiment classification models applied to the 22,641 cleaned reviews, the customer sentiment is overwhelmingly positive.

- **TextBlob Analysis:**
  - **Positive:** 94.1% (21,300 reviews)
  - **Negative:** 5.2% (1,174 reviews)
  - **Neutral:** 0.7% (167 reviews)
- **VADER Analysis:**
  - **Positive:** 96.5% (21,853 reviews)
  - **Negative:** 2.4% (549 reviews)
  - **Neutral:** 1.1% (239 reviews)

Insight: Both models indicate that over 94% of customers report a positive sentiment regarding their purchases.

Most Frequent Words in Reviews
The text visualization and tokenization processes highlighted specific keywords associated with different sentiments.

Positive Reviews:
- **Descriptors:** "Love", "Great", "Perfect", "Fit", "Flatter", "Like".
- **Context:** Customers frequently praise the "fit" and the "look" of the items.
- **Items:** "Dress" and "Top" are the most frequently mentioned items in positive contexts.

Negative Reviews:
- **Descriptors:** "Small", "Big", "Little", "Return", "Disappointed".
- **Specific Pain Points:**
  - **"Fabric":** Often associated with poor quality or texture.
  - **"Zipper":** Suggests functional hardware failures.
  - **"Back":** Likely referring to fit issues in the back of garments.
- **Context:** The word "Dress" appears frequently here as well, indicating it is a high-risk category for returns or fit issues.

Key Products with Most Positive Feedback
The analysis aggregated high-rated reviews (4 and 5 stars) by department to identify top performers.

- **Highest Volume (Most Popular):**

    - **Tops:** 7,652 high ratings (5,444 five-star reviews). This is the primary driver of positive sentiment volume.
    - **Dresses:** 4,634 high ratings. Despite some negative keywords appearing in the text analysis, dresses remain a massive volume driver for positive engagement.
- **Highest Satisfaction (Best Average Rating):**
    - **Jackets:** Average rating of **4.76/5.0**. While lower in volume (804 reviews), this category has the highest customer satisfaction rate.
    - **Bottoms:** Average rating of **4.74/5.0**.
    - **Intimates:** Average rating of **4.74/5.0**.

Business Recommendations

Based on the sentiment data and category performance:

1. **Address Sizing and Fit Issues:**
    - Negative reviews frequently cite "small," "big," and "fit." The business should audit size charts for accuracy, particularly for **Dresses**, which appear in both positive and negative frequent word lists. Implementing a "true-to-size" customer feedback loop on product pages could reduce returns.
2. **Quality Control of Hardware and Fabric:**
    - Keywords like "zipper," "fabric," and "cheap" in negative reviews suggest specific manufacturing defects. A targeted quality control check on zipper suppliers and fabric density could significantly reduce the 5.8% negative sentiment rate.
3. **Capitalize on "Jackets" and "Tops":**
    - **Tops** are the volume leader; marketing spend should continue here to acquire new customers.
    - **Jackets** have the highest satisfaction; the business should consider expanding this product line or cross-selling jackets to customers buying bottoms, given the high trust in this category.
4. **Automated Sentiment Alerting:**
    - Since 93%+ of reviews are positive, the business should implement an automated alert system (using the VADER logic demonstrated in the notebook) to flag the rare <0.05 sentiment scores for immediate customer service intervention, turning potential detractors into promoters.