# Progress Report: Customer Review Mining in Beauty & Personal Care

## 1. Introduction

The objective of this project is to analyze customer reviews in the Beauty & Personal Care sector, focusing on five selected products/brands: Maybelline SuperStay Matte Ink Lipstick, CeraVe Moisturizing Cream, Neutrogena Hydro Boost Water Gel, Drunk Elephant Protini Polypeptide Cream, and Tarte Shape Tape Concealer. The project applies Natural Language Processing (NLP) techniques to extract insights into customer sentiment, common themes, and overall product perceptions.

## 2. Data Collection

Initially, the project attempted to scrape reviews directly from Amazon using Selenium and Python. However, due to Amazon's strong anti-bot protections (CAPTCHAs and blocked requests), the scraping approach was not feasible. As a result, the project pivoted to a more reliable data source: the Amazon Beauty Reviews Dataset from Kaggle. This dataset contains hundreds of thousands of reviews with structured fields such as ratings, review text, and product metadata.

## 3. Methods and Technologies Used

- Data Handling: Python (Pandas, NumPy)
- Natural Language Processing (NLP): VADER Sentiment Analysis, Text Preprocessing (tokenization, cleaning)
- Visualization: Matplotlib for initial charts, Tableau planned for interactive dashboards
- Environment: Jupyter Notebook for iterative analysis
- Dataset: Amazon Beauty Reviews Dataset (Kaggle)

## 4. Initial Results

A filtered dataset of 347 reviews was created, focusing on the five chosen brands. The following early findings were observed:

- Rating Distribution: Majority of reviews are in the 4-star and 5-star categories, indicating positive product reception.
- Sentiment Analysis: Using the VADER sentiment analyzer, reviews were classified as Positive, Negative, or Neutral. From the sample:
  - Positive: 314 reviews (~88%)
  - Negative: 22 reviews (~6%)
  - Neutral: 11 reviews (~3%)

These results align closely with star ratings, confirming the dataset's reliability for sentiment modeling.

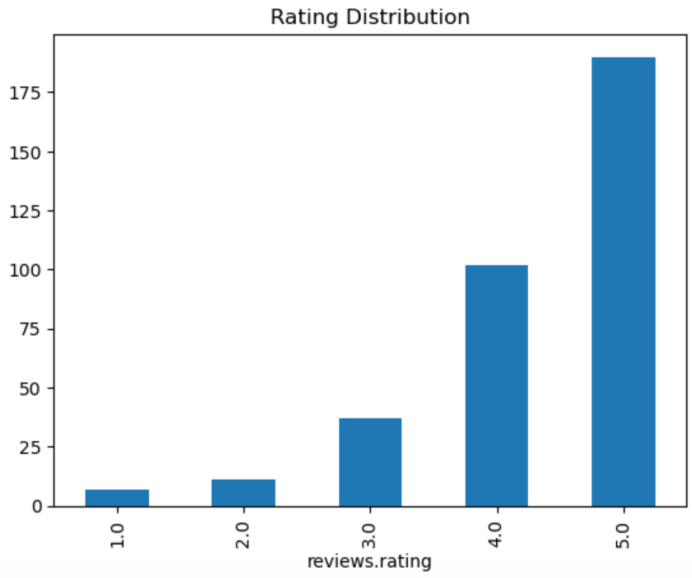
[7]:
```
brands = ["Maybelline", "CeraVe", "Neutrogena", "Drunk Elephant", "Tarte"]

filtered_df = df[df['reviews.text'].str.contains('|'.join(brands), case=False, na=False)]
print(filtered_df.shape)
filtered_df[['reviews.rating','reviews.title','reviews.text']].head()
```

```
(347, 21)
```

[7]:

| | reviews.rating | reviews.title | reviews.text |
|---|---|---|---|
| **408** | 4.0 | Good for the price | It's a good starter tablet, to bad it runs on ... |
| **431** | 3.0 | Ok for the price | Good starter unit. Easy for a beginner to use.... |
| **436** | 5.0 | Great little online browser for the price | I bought these as a starter online browser. It... |
| **1057** | 5.0 | Purchased for Mother as a present | And she has been happy about it ever since. No... |
| **1207** | 3.0 | Not ad happy with this one by far | I have had every version of the Kindle since t... |

[11]:
```
<Axes: title={'center': 'Rating Distribution'}, xlabel='reviews.rating'>
```




[15]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()

def label_sentiment(text):
    s = analyzer.polarity_scores(str(text))['compound']
    if s >= 0.05: return 'positive'
    if s <= -0.05: return 'negative'
    return 'neutral'

filtered_df['sentiment'] = filtered_df['reviews.text'].map(label_sentiment)
filtered_df['sentiment'].value_counts()
```

```
/var/folders/hw/z2tsjsz17wx21b2my8h9zzb00000gn/T/ipykernel_35706/2296251034.py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  filtered_df['sentiment'] = filtered_df['reviews.text'].map(label_sentiment)
```

[15]:
```
sentiment
positive    314
negative     22
neutral      11
Name: count, dtype: int64
```

## 5. Next Steps

- Perform topic modeling (e.g., LDA or BERTopic) to identify recurring themes such as hydration, smudge-proof quality, irritation, and packaging.
- Build interactive dashboards in Tableau or Power BI for brand-wise and product-wise comparison.
- Generate actionable insights highlighting customer likes and dislikes to inform product improvements and marketing strategies.
- Finalize documentation and prepare a comprehensive research report.