Sentiment Analysis Report

Dataset Description

The dataset used for this project is the **Amazon Consumer Reviews dataset** provided by Datafiniti. It contains customer product reviews, including textual feedback in the column **"reviews.text"**, which serves as the primary feature variable for this analysis. These reviews reflect customer opinions on various Amazon products and form the basis for sentiment analysis.

Preprocessing Steps

Before applying sentiment analysis, the review text underwent several preprocessing steps to improve accuracy:

- **Lowercasing**: All text was converted to lowercase for consistency.
- **Whitespace trimming**: Extra spaces were removed to standardize text.
- **Stopword removal**: Common words such as *"the"*, *"is"*, and *"of"* were removed, as they add little value to sentiment classification.
- **Filtering tokens**: Only alphabetic tokens were retained, ensuring that symbols, numbers, and irrelevant characters did not interfere with the analysis.

These steps produced a cleaned dataset of reviews better suited for meaningful sentiment classification.

Evaluation of Results

The sentiment analysis model, implemented with **spaCy** and **spacytextblob**, was applied to the cleaned dataset. Each review was classified as **Positive**, **Negative**, or **Neutral**, based on its polarity score ranging from -1 (very negative) to +1 (very positive).

The majority of the reviews were classified as **Positive**, indicating that most customers expressed satisfaction with their purchases. A smaller portion of reviews were classified as Negative or Neutral, showing that while some customers had concerns or mixed feelings, overall sentiment leaned strongly positive.

Insights and Limitations

Strengths:

- The model provides a quick and accessible way to categorize customer reviews.
- Businesses can use these insights to gauge overall customer satisfaction and track trends in product feedback.
- The preprocessing pipeline ensures that irrelevant text elements do not affect the results.

Limitations:

- The model struggles with **sarcasm, mixed opinions, or subtle dissatisfaction**. For example, a review such as *"The product works, but I expected more"* may be classified as Neutral despite expressing disappointment.
- Sentiment strength is based only on polarity, which may oversimplify complex emotions in text.
- More advanced models such as **BERT** or transformer-based approaches could provide greater accuracy by understanding context and nuanced expressions.