Abusufyan Jamadar
AJ23110011464
**Skills Bootcamp in Data Science (Fundamentals)**

# Sentiment Analysis Report

1. Description of the Dataset Used:

The datasets utilized for this sentiment analysis project are extracted from Amazon Consumer Reviews, also known as 'Amazon Reviews Dataset'. These datasets contain product reviews written by Amazon customers for various products. The datasets include information such as the text of each review, the title of the review, the star rating, and other metadata related to products. Specifically, the analysis was conducted on three separate files:

- `1429_1.csv`
- `Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv`
- `Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv`

Each dataset contains reviews across different product categories, providing a diverse set of texts for sentiment analysis.

2. Details of the Preprocessing Steps:

The preprocessing of text data is a crucial step before performing sentiment analysis. The preprocessing steps implemented in this project include:

- Text Normalization: Converting all text to lower case to ensure that the algorithm treats words like "The" and "the" equally.
- Tokenization and Lemmatization: Splitting the text into individual words or tokens and then converting each token to its base or dictionary form (lemma).
- Stopwords Removal: Eliminating common words that do not contribute significantly to the sentiment of the text, such as "is", "and", "the", etc.
- Punctuation Removal: Removing all punctuation marks, as they are not necessary for sentiment analysis.
- Batch Processing: Processing texts in batches for more efficient analysis.

These steps were applied to the 'reviews.text' column of each dataset, resulting in a cleaned and preprocessed version of the review texts.

3. Evaluation of Results:

The sentiment analysis was performed using the `spacytextblob` extension for the spaCy library. Each review was classified as 'Positive', 'Negative', or 'Neutral' based on the polarity score obtained from `spacytextblob`. In the sample analysis conducted:

- All five sample reviews from the first dataset were classified as 'Positive'.

- The effectiveness of the model was observed through these classifications based on the context provided by each review.

However, the actual sentiment analysis may require further validation and testing against a benchmark or labeled dataset to determine its accuracy and reliability systematically.

4. Insights into the Model's Strengths and Limitations:

Strengths:

- The model provides a quick and automated way to gauge the sentiment of large volumes of text.

- It can process and analyze text data in batches, enhancing efficiency.

- The model is capable of identifying the general sentiment of texts, aiding in understanding consumer sentiments at scale.

Limitations:

- The `en_core_web_sm` spaCy model used does not inherently support sentiment analysis. Integration with `spacytextblob` is required for sentiment evaluation.

- The sentiment analysis may not always align with human judgment due to the complexity and subtlety of human emotions expressed in text.

- The model's performance heavily depends on the quality and range of the training data used by `spacytextblob`.

- Similarity measures between reviews might not provide useful insights without substantial contextual understanding and word vectors.

Future work could involve enhancing the sentiment analysis model with more sophisticated NLP techniques and training it on a dataset specifically labeled for sentiment to improve its accuracy and reliability. Additionally, exploring larger spaCy models or different sentiment analysis libraries could yield better results.