Capstone Project – NLP Sentiment Analysis

Objective of the Project:

The objective of this capstone project is to apply Natural Language Processing (NLP) techniques to analyze customer sentiment in product reviews. Specifically, the project aims to:

1. Preprocess textual data from Amazon product reviews, including cleaning, lowercasing, and removing stop words.
2. Perform sentiment analysis using the spaCy NLP library combined with spaCyTextBlob to classify reviews as positive, negative, or neutral.
3. Evaluate and summarize sentiment patterns in the dataset to understand overall customer satisfaction.
4. Generate insights and recommendations based on sentiment trends, highlighting strengths and limitations of the model.

The broader goal is to demonstrate practical NLP skills in processing real-world textual data and producing actionable insights, which can be leveraged by businesses to improve products, services, or customer experience.

---

## 1. Dataset Description

- **Source:** Amazon product reviews dataset collated by Datafiniti (Kaggle).
- **Size:** 34,659 reviews after removing missing values.
- **Columns of interest:**
  - reviews.text: The product review content used for sentiment analysis.
  - reviews.rating: Original star rating (1–5).

The dataset contains reviews for multiple Amazon products, primarily electronics like tablets and Kindle devices.

---

## 2. Preprocessing Steps

1. **Lowercasing:** All text converted to lowercase to standardize data.
2. **Whitespace stripping:** Leading/trailing spaces removed.
3. **Stop word removal:** Words like “the”, “is”, “of” removed using spaCy.
4. **Alphabetic filtering:** Only alphabetic tokens retained (punctuation and numbers removed).

**Sample preprocessed reviews:**

- "awesome product backlit shows page number easier turn pages"
- "love easy learn operate quick response easy hook"

---

### 3. Sentiment Analysis Method

- **Library:** spaCy + spaCyTextBlob
- **Polarity Scores:**
    - Range: -1 (very negative) to +1 (very positive)
    - Determined sentiment label:
        - Positive: polarity > 0
        - Neutral: polarity = 0
        - Negative: polarity < 0
- **Example Sentiment Results:**

    | Review | Polarity | Sentiment |
    |--------|----------|-----------|
    | awesome product backlit ... | 1.0 | Positive |
    | screen adequate service ... | 0.042 | Positive |
    | product helpful excellent ... | 0.4 | Positive |

### 4. Similarity Analysis

- Similarity between first two reviews: **0.68**
- Indicates reviews with overlapping words or similar content can be identified.
- Useful for detecting duplicate or highly similar reviews.

### 5. Overall Sentiment Distribution

| Sentiment | Count | Percentage |
|---|---|---|
| Positive | 30,347 | 87% |
| Neutral | 2,809 | 8% |
| Negative | 1,503 | 4% |

**Insights:**

- Majority of reviews are positive, suggesting high customer satisfaction.
- Neutral and negative reviews are low, indicating generally favorable product reception.

### 6. Strengths of the Model

- Able to distinguish positive, neutral, and negative sentiments.
- Preprocessing ensures consistent and clean input for analysis.

### 7. Limitations of the Model

- spaCyTextBlob sentiment scores are **lexicon-based**, so very nuanced sentiment (e.g., sarcasm) may not be detected.
- Polarity scores may not reflect intensity in very long reviews.
- Context-specific meanings (e.g., “light” as weight vs. brightness) may not always be correctly interpreted.

---

### 8. Conclusion

- The sentiment analysis pipeline successfully processed **all 34,659 reviews**, providing polarity scores and sentiment labels.
- Positive reviews dominate, highlighting overall user satisfaction.
- The processed CSV (amazon_reviews_sentiment.csv) can be used for further reporting or visualization.