5.1 Dataset Description

The dataset used for this analysis contains product reviews from amazon. The relevant column for our analysis is labelled 'reviews.text'.

5.2 Preprocessing Steps

1. **Loading the Model and Dataset:**
    - A spaCy model with word vectors (e.g., en_core_web_lg) is loaded.
    - The reviews dataset is read from amazon_product_reviews.csv'.
2. **Data Cleaning:**
    - Missing values in the 'reviews.text' column are removed, resulting in a cleaned df named clean_data.

5.3 Evaluation of Results

1. **Sentiment Analysis:**
    - The TextBlob library is used to predict sentiment, positive or negative, for product reviews.
    - An example review (“This product is amazing!”) is analysed, and the predicted sentiment is printed.
2. **Similarity Score Calculation:**
    - The spaCy model calculates the similarity score between the first two product reviews (review1 and review2).
    - The similarity score ranges from 0 to 1, where higher values indicate greater similarity.

5.4 Insights into the model's strengths and limitations

- **Strengths:**
    - The use of spaCy and TextBlob allows for efficient sentiment analysis and similarity scoring.
    - The model can handle large volumes of text data.
- **Limitations:**
    - The dataset’s origin and context are unspecified.
    - We lack information on the specific product category or domain.
    - The model’s performance may vary based on the quality and diversity of reviews.