This is a 50k data sets evenly distributed dataset 25k are in Positive review and 25k are negative.Only 2 feature is in this data set one is review and other one is sentiment. Every review has roughly 150 to 250 words.All reviews are in English language and the sentiment categorize as “positive” and “negative” .There are no missing or null reviews I data set.There are 49582 unique values in the dataset.

**1. Cleaning the Text:**

* **Removing HTML tags:** If your reviews contain HTML tags, you can use libraries like BeautifulSoup to remove them.
* **Removing punctuation and special characters:** Punctuation and special characters usually don't contribute much to sentiment analysis and can be removed using regular expressions.
* **Converting to lowercase:** Converting all text to lowercase ensures that words are treated the same regardless o 1. Cleaning the Text:
* Removing HTML tags: If your reviews contain HTML tags, you can use libraries like BeautifulSoup to remove them.
* Removing punctuation and special characters: Punctuation and special characters usually don't contribute much to sentiment analysis and can be removed using regular expressions.
* Converting to lowercase: Converting all text to lowercase ensures that words are treated the same regardless of their capitalization.
* Removing stop words: Stop words (e.g., "the," "a," "is") are common words that don't carry much meaning and can be removed using libraries like NLTK.
* f their capitalization.
* **Removing stop words:** Stop words (e.g., "the," "a," "is") are common words that don't carry much meaning and can be removed using libraries like NLTK.