Data: Our data will be sourced from Kaggle, specifically Amazon Product Reviews. The dataset exists in a CSV format, which we will interpret and manipulate using the Pandas library. In this dataset, we are most concerned with the “score”, “summary”, and “text” columns. The “score” solemn contains categorical values ranging from 1 to 5 describing how satisfied the customer is with the product; 5 being satisfied and 1 being unsatisfied. The “summary” column contains a short, user defined text summarizing the content of their product review. Finally, the “text” column is the review itself.

In this project, the cleaning process began by removing punctuation marks using a custom function that filtered out any characters found in Python’s string.punctuation set. Next, numbers embedded within the text were removed using regular expressions, followed by the removal of accented characters through Unicode normalization, which converts accented letters to their ASCII equivalents. Additionally, special characters and non-alphanumeric symbols were eliminated using another regular expression pattern, further simplifying the text.