NAME: Krushal Kalkani
Z-NO: Z23808174

SUBJECT: Intro to Data Science (CAP 5768)

TASK 1 - PROBLEM SELECTION, DATA COLLECTION, AND DESCRIPTION

Sentiment analysis has become a critical tool for businesses in understanding public perception and customer sentiment from large volumes of unstructured text data. With the rise of online platforms, consumers frequently share their experiences and opinions, making it essential for companies to gauge customer satisfaction and feedback effectively. This project aims to address the challenge of automating the sentiment analysis process on customer reviews, allowing businesses to quickly classify reviews as positive, negative, or neutral.

The goal is to build a comprehensive end-to-end sentiment analysis pipeline using Python. By exploring various NLP techniques and machine learning algorithms, the model will be able to accurately classify sentiment in text, providing actionable insights for businesses to enhance their products and services. Key preprocessing steps and feature extraction methods will be employed to refine data and improve model performance.

Dataset Overview:

- **Size**: Approximately 3,000,000 records
- **Type**: Text-based CSV (Tabular)
- **Source**: Kaggle (https://registry.opendata.aws/amazon-reviews/) or UCI Machine Learning Repository
- **Number of Variables**: 5

1. **Review_Text**: Main content of each customer review.
2. **Star_Rating**: Ratings from 1 to 5, which can be mapped to sentiment labels (e.g., 1-2 for negative, 3 for neutral, 4-5 for positive).
3. **Product_ID**: Identifier for the product reviewed.
4. **Product_Category**: Category of the product reviewed (e.g., electronics, books, apparel).
5. **Review_Date**: Date of the review submission