#data preprocessing

fake reviews dataset.csv

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

nltk.download('stopwords')

nltk.download('punkt')

# Load Dataset (Replace with your dataset path or URL)

df = pd.read\_csv("https://raw.githubusercontent.com/julienc91/labeled-amazon-reviews/master/data/labeled\_data.csv")

# Display first few rows

print(df.head())

# Check for missing values

print(df.isnull().sum())

# Basic statistics

print(df.describe())

# Plot review distribution

sns.countplot(x=df['label']) # Assuming 'label' column has fake/real labels

plt.title("Distribution of Fake vs. Real Reviews")

plt.show()

# Text Preprocessing

def preprocess\_text(text):

text = text.lower() # Convert to lowercase

tokens = word\_tokenize(text) # Tokenization

tokens = [word for word in tokens if word.isalnum()] # Remove special characters

tokens = [word for word in tokens if word not in stopwords.words('english')] # Remove stopwords

return ' '.join(tokens)

# Apply preprocessing

df['cleaned\_review'] = df['review\_text'].apply(preprocess\_text) # Replace 'review\_text' with the actual column name

print(df[['review\_text', 'cleaned\_review']].head())