**DATASETS IDENTIFICATION-ASSIGNMENT1**

**Dataset Explanation:**

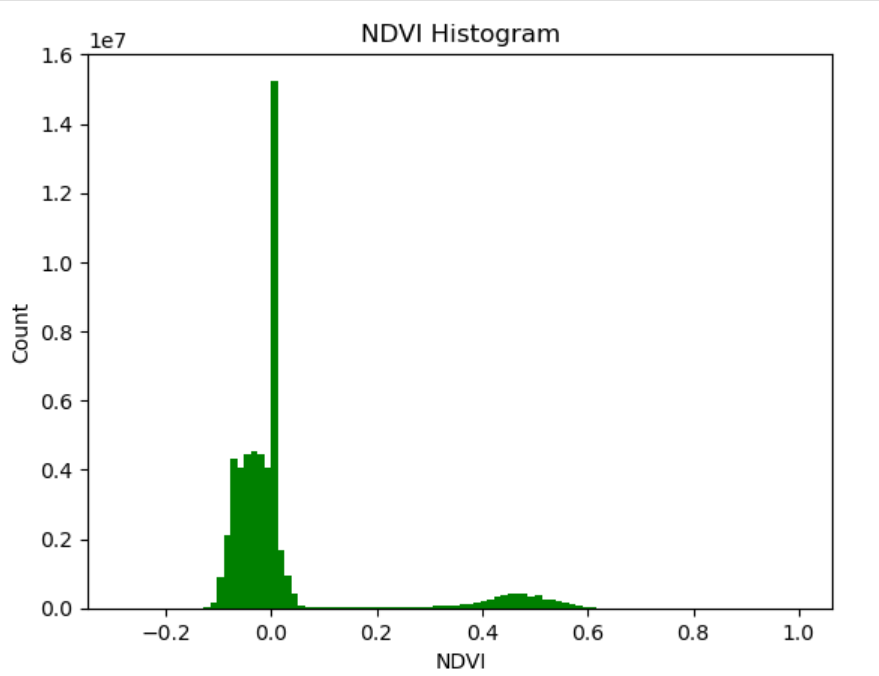
**IMAGE DATASET – LISS-III Satellite Imagery (Resourcesat-2)**

* **Source:** ISRO Bhoonidhi Open Data Portal
* **Sensor:** LISS-III (Linear Imaging Self Scanner) onboard Resourcesat-2
* **Spatial Resolution:** 23.5 meters
* **Bands Used:**
  + **BAND2 – Green (0.52–0.59 µm)**
  + **BAND3 – Red (0.62–0.68 µm)**
  + **BAND4 – Near Infrared (0.77–0.86 µm)**
  + **BAND5 – Short-Wave Infrared (1.55–1.70 µm)**
* **File Format:** GeoTIFF (multiband remote sensing raster images)
* **Use Cases:**
  + Vegetation analysis (NDVI)
  + Land cover classification
  + Water body detection
  + Deforestation monitoring
* **Why it’s useful:** These satellite bands enable visual and spectral analysis for real-world environmental insights.

**TEXT DATASET – Bangla Online Comments Dataset**

* **Source:** Mendeley Data (Bangla Online Comments v1)
* **Format:** Excel .xlsx
* **Total Comments:** 44,001
* **Languages:** Bengali (Bangla script)
* **Labelled Classes:**
  + Non-Bully
  + Sexual
  + Threat
  + Troll
  + Religious
* **Metadata Included:** Gender, profession of victim, engagement stats
* **Use Cases:**
  + Hate speech & bullying detection
  + Text classification
  + Demographic bias analysis
* **Why it’s useful:** It provides a real-world labeled dataset in a low-resource language, enabling deep NLP tasks in regional contexts.

**EDA for image dataset:**



**EXPLANATION/INTERPRETATION:**

I have used ISRO LISS-III data bands (B3: Red, B4: NIR) to compute NDVI for vegetation health assessment. The histogram of NDVI shows two major regions — low values (~0) for water or barren land and moderate positive values (0.2–0.6) for healthy vegetation. This confirms spatial land cover variation and supports use in environmental monitoring.

**CODE :**

import zipfile

with zipfile.ZipFile("R2301FEB2025071554009600063PSANSTUC00GTDF.zip", 'r') as zip\_ref:

zip\_ref.extractall("liss3\_data")

!pip install rasterio

import rasterio

import numpy as np

import matplotlib.pyplot as plt

import os, glob

folder = "liss3\_data/R2301FEB2025071554009600063PSANSTUC00GTDF"

bands = {}

for f in glob.glob(folder + "/\*.tif"):

band = os.path.basename(f).replace(".tif", "")

with rasterio.open(f) as src:

bands[band] = src.read(1).astype("float32")

nir = bands['BAND4']

red = bands['BAND3']

green = bands['BAND2']

ndvi = (nir - red) / (nir + red + 1e-5)

plt.hist(ndvi.flatten(), bins=100, color="green")

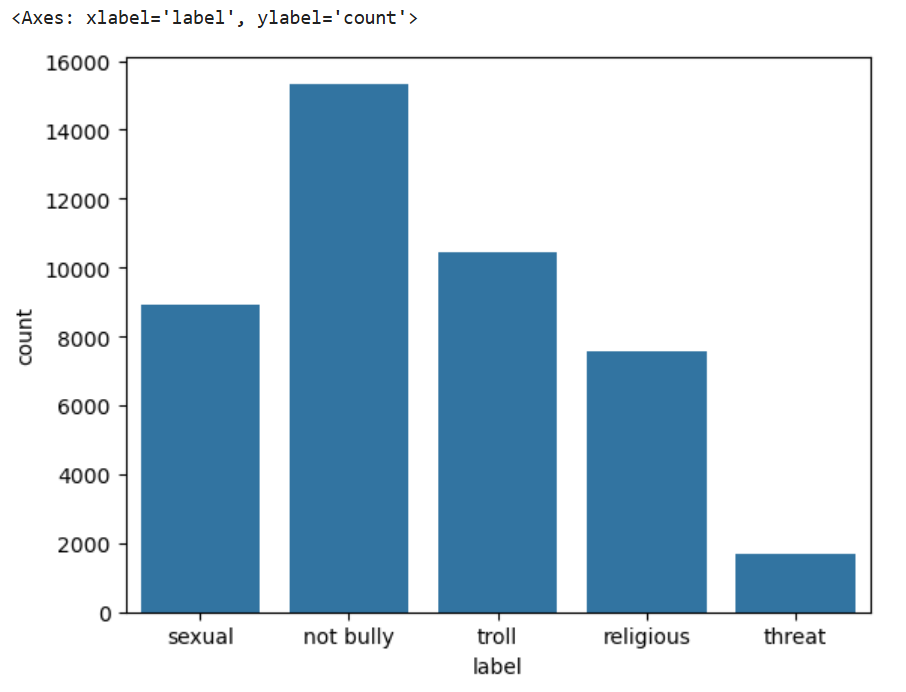
plt.title("NDVI Histogram")

plt.xlabel("NDVI")

plt.ylabel("Count")

plt.show()

**EDA for Text dataset:**



**EXPLANATION/INTERPRETATION:**

The dataset contains Bangla text comments with labels such as ‘not bully’, ‘sexual’, ‘troll’, etc. EDA revealed a clear class imbalance, with most samples labeled as ‘not bully’. I have also analyzed comment lengths and label counts, which can help in designing better preprocessing and classification strategies.

**CODE:**

import pandas as pd

import seaborn as sns

df = pd.read\_excel("bangla\_online\_comments\_dataset.xlsx")

sns.countplot(x='label', data=df)