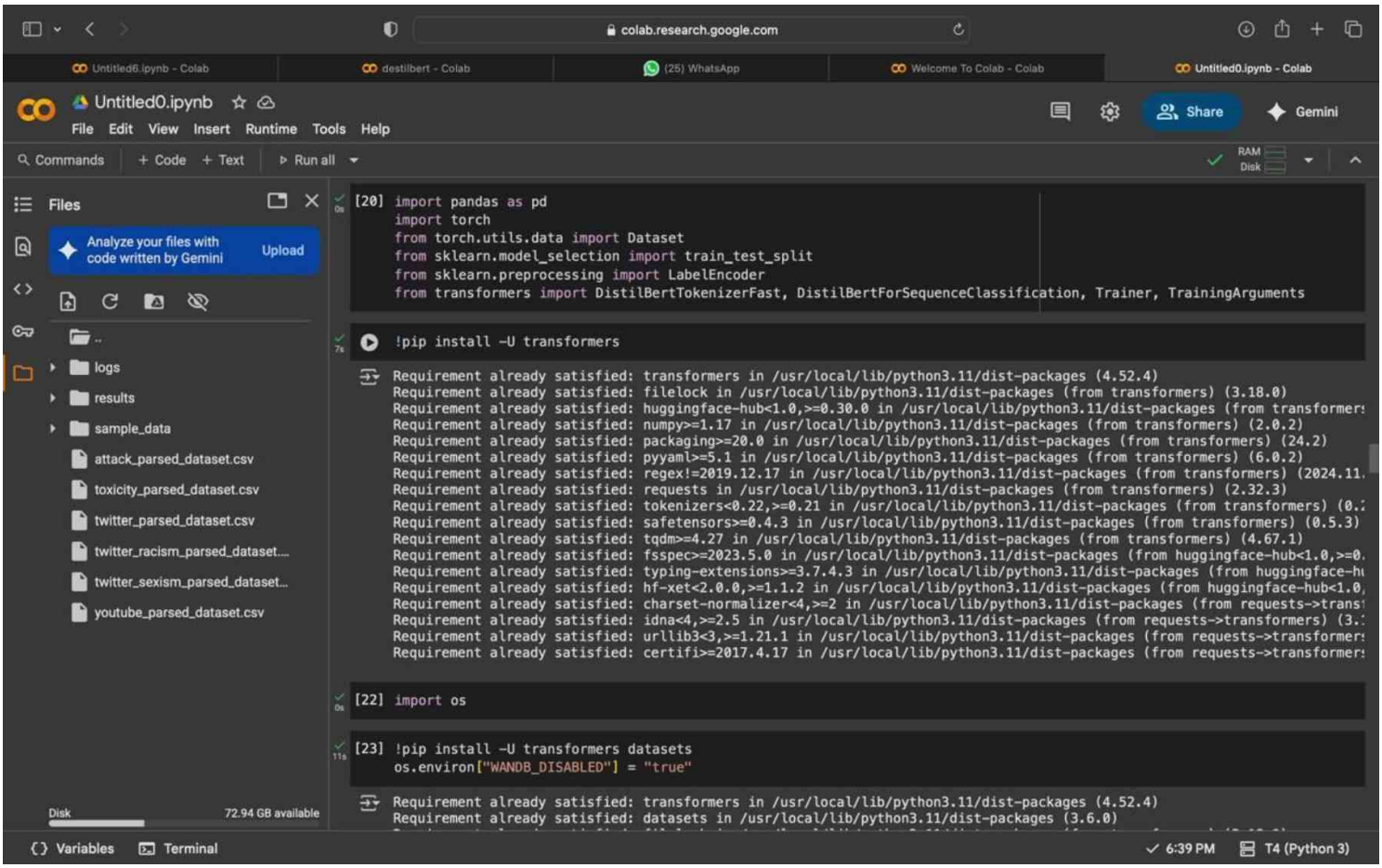
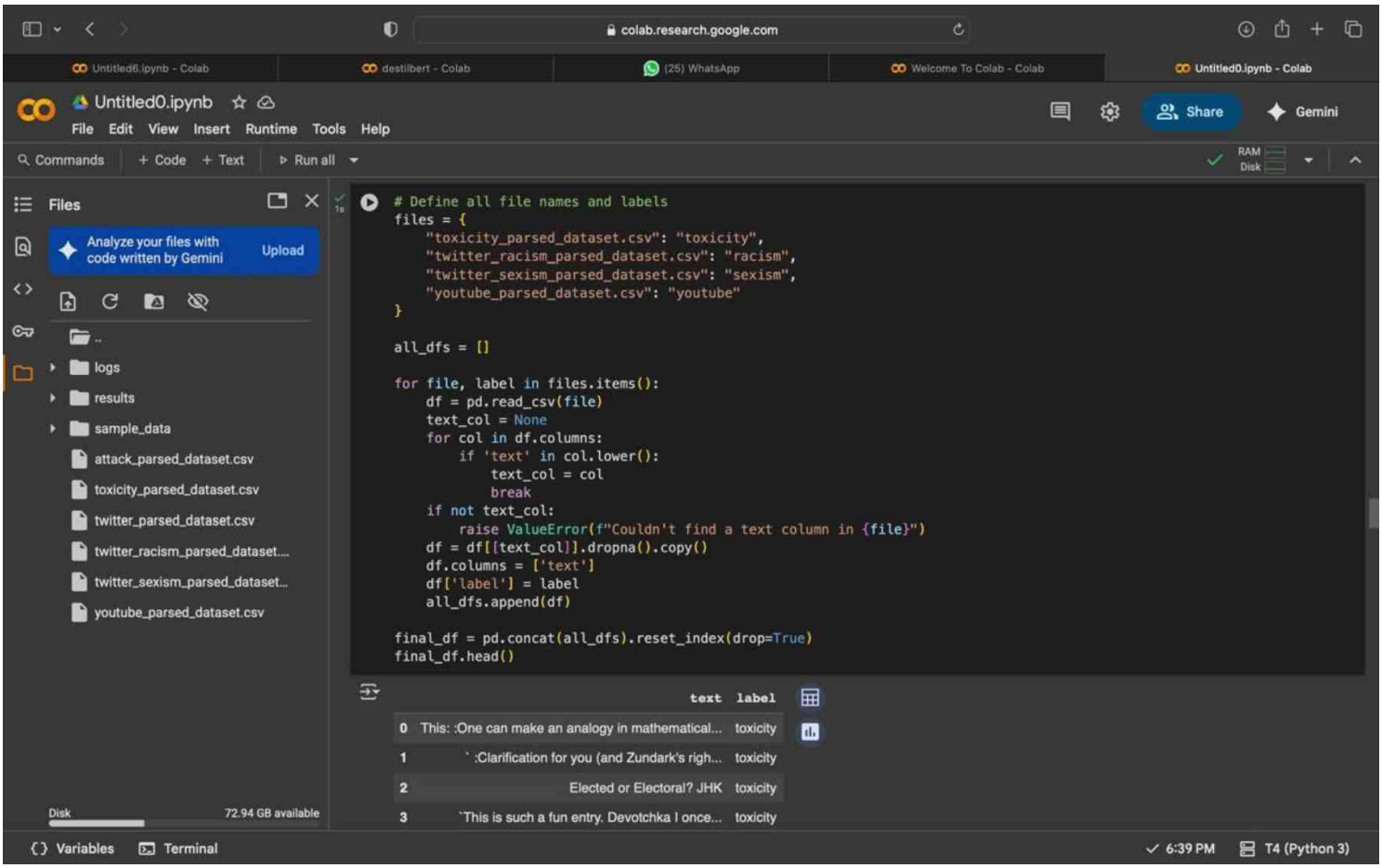
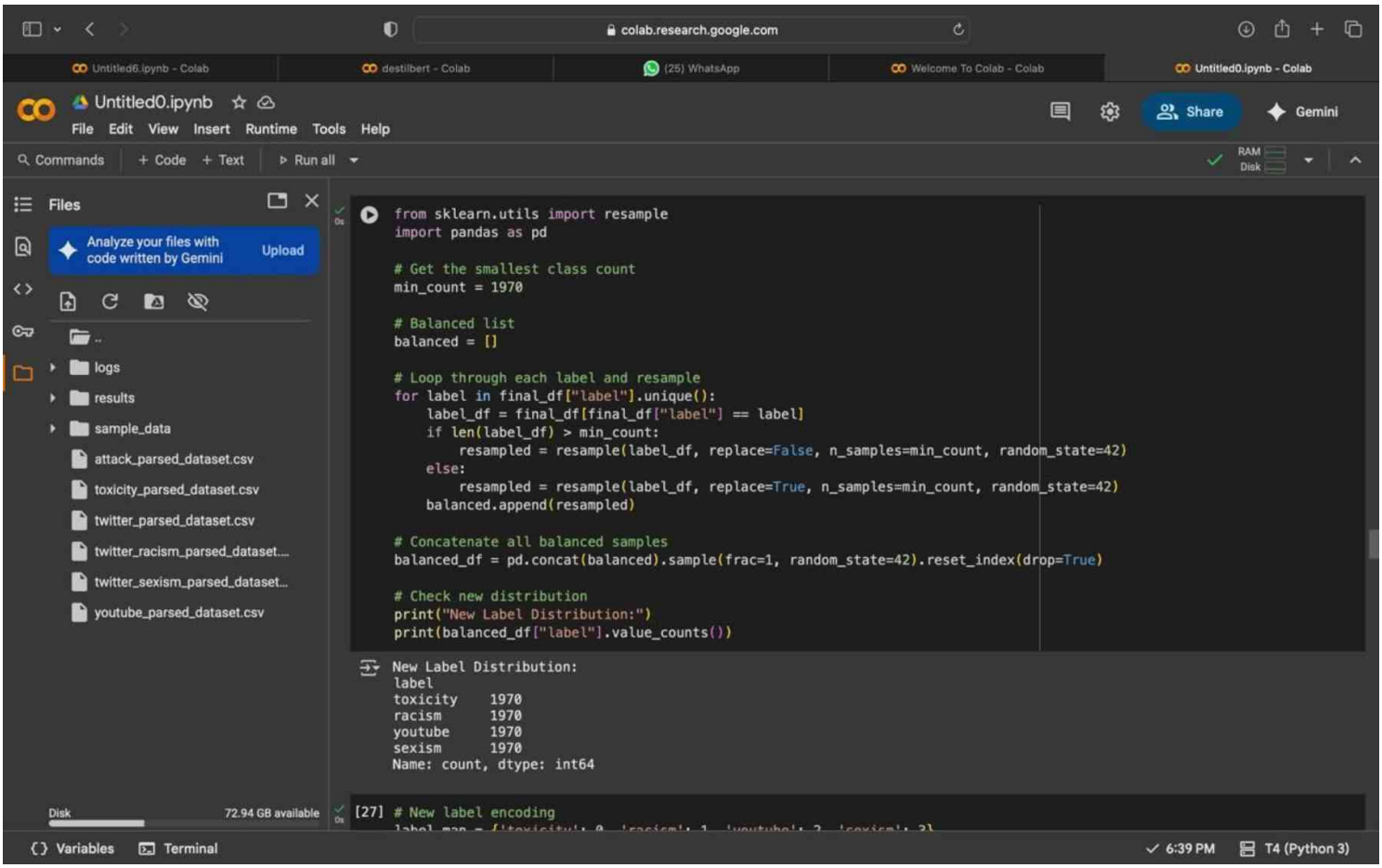
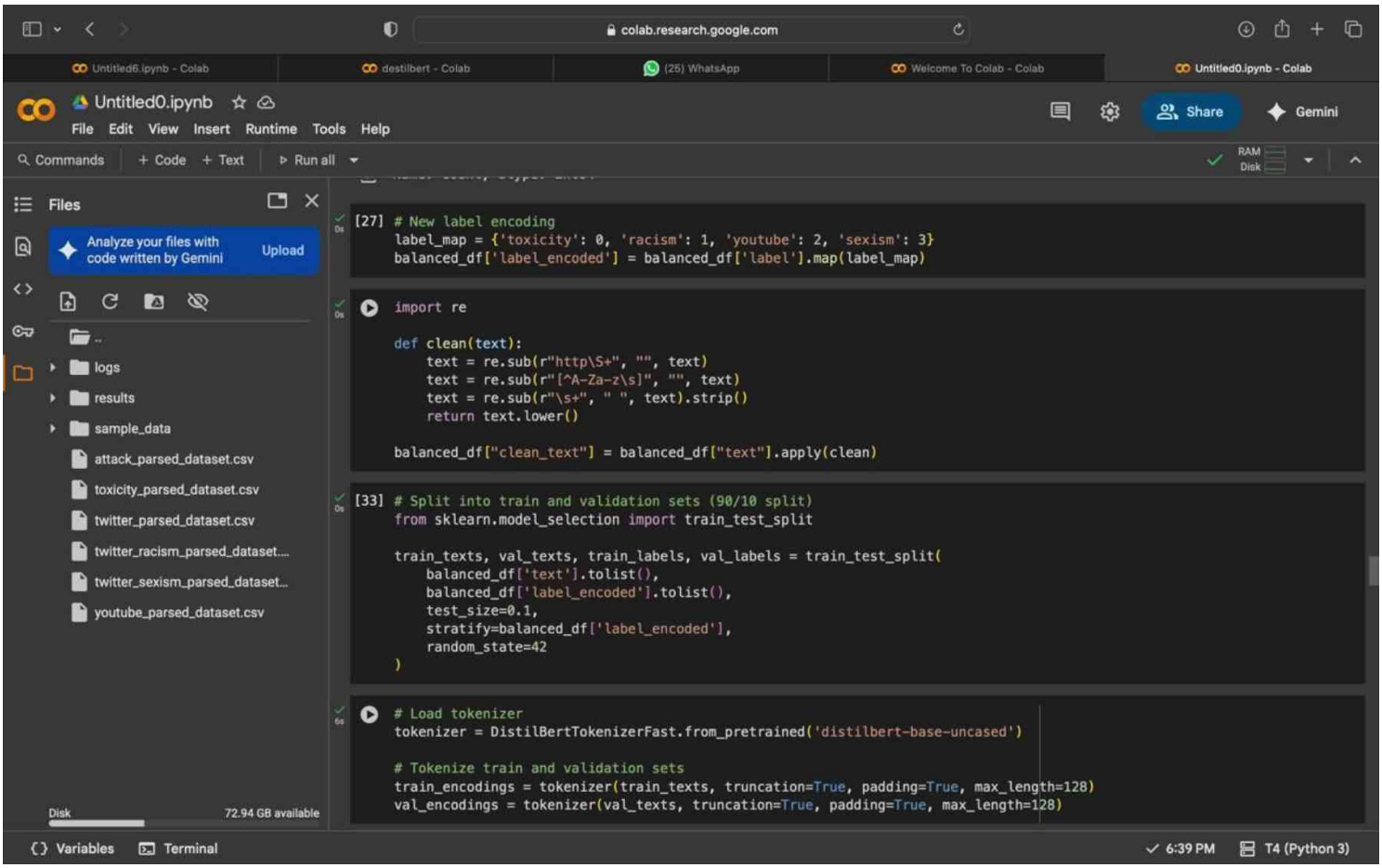
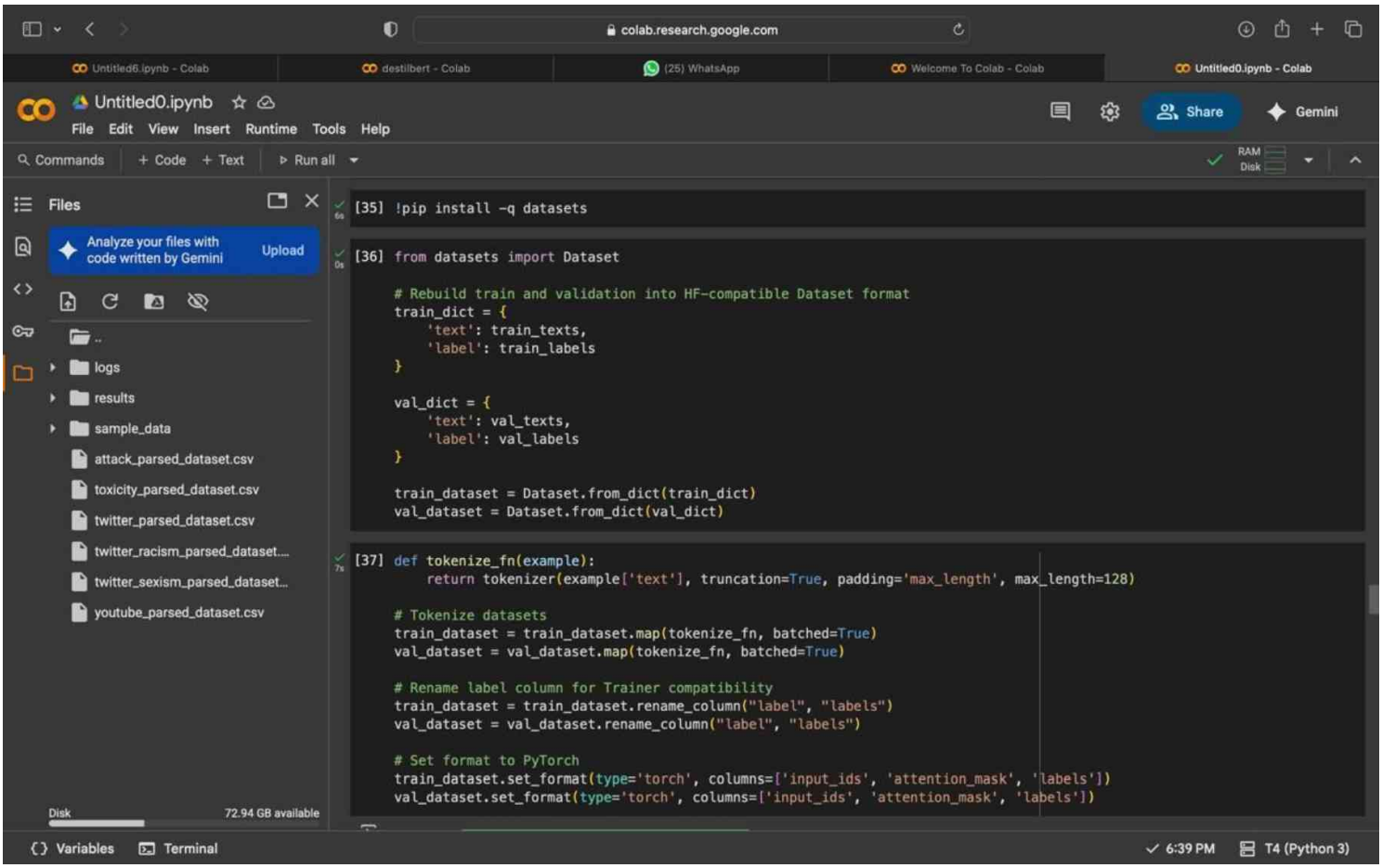
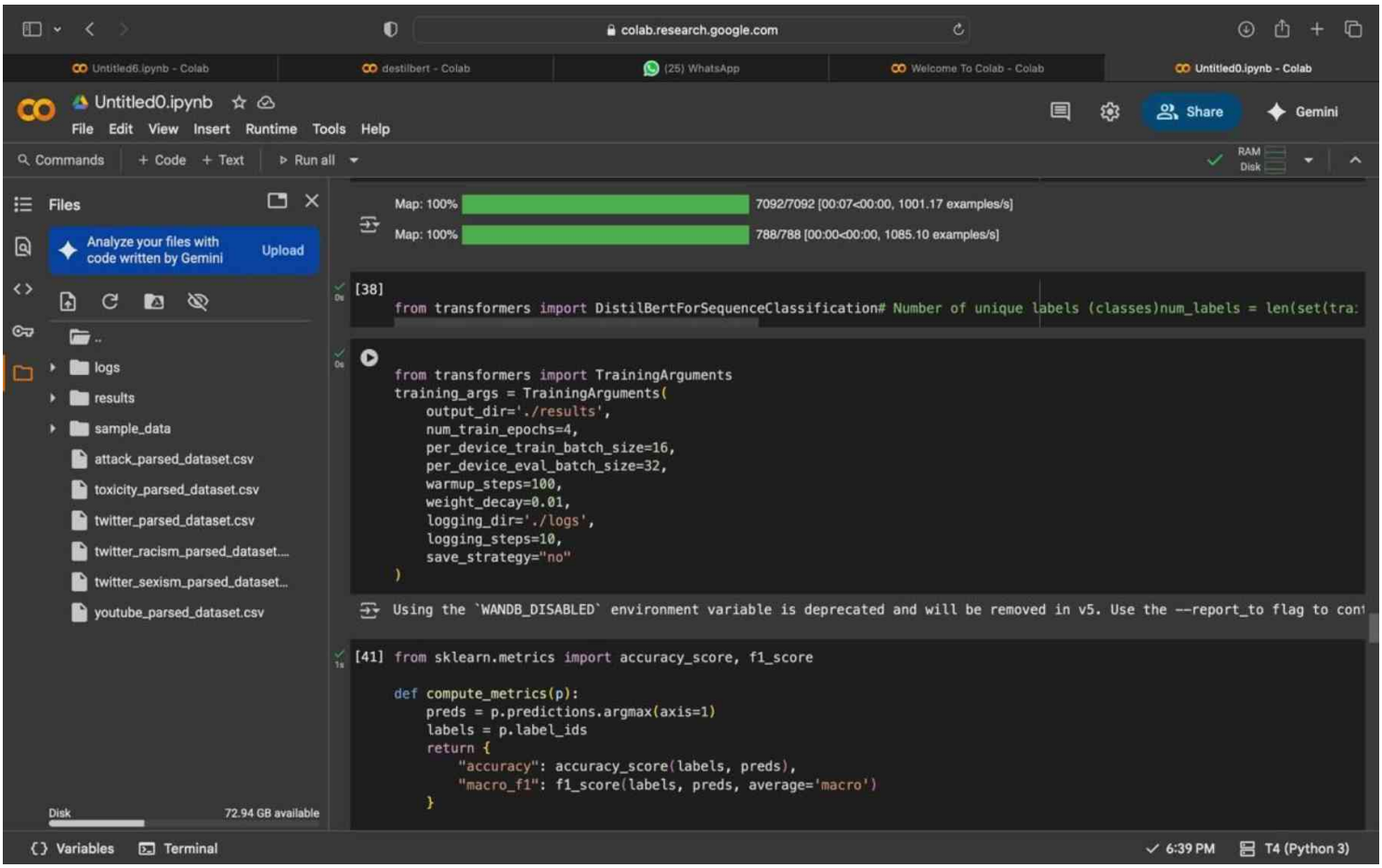

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk 72.94 GB available
[20] import pandas as pd
import torch
from torch.utils.data import Dataset
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from transformers import DistilBertTokenizerFast, DistilBertForSequenceClassification, Trainer, TrainingArguments
!pip install -U transformers
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hu
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transf
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers
[22] import os
[23] !pip install -U transformers datasets
os.environ["WANDB_DISABLED"] = "true"
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (3.6.0)
Variables Terminal
6:39 PM T4 (Python 3)

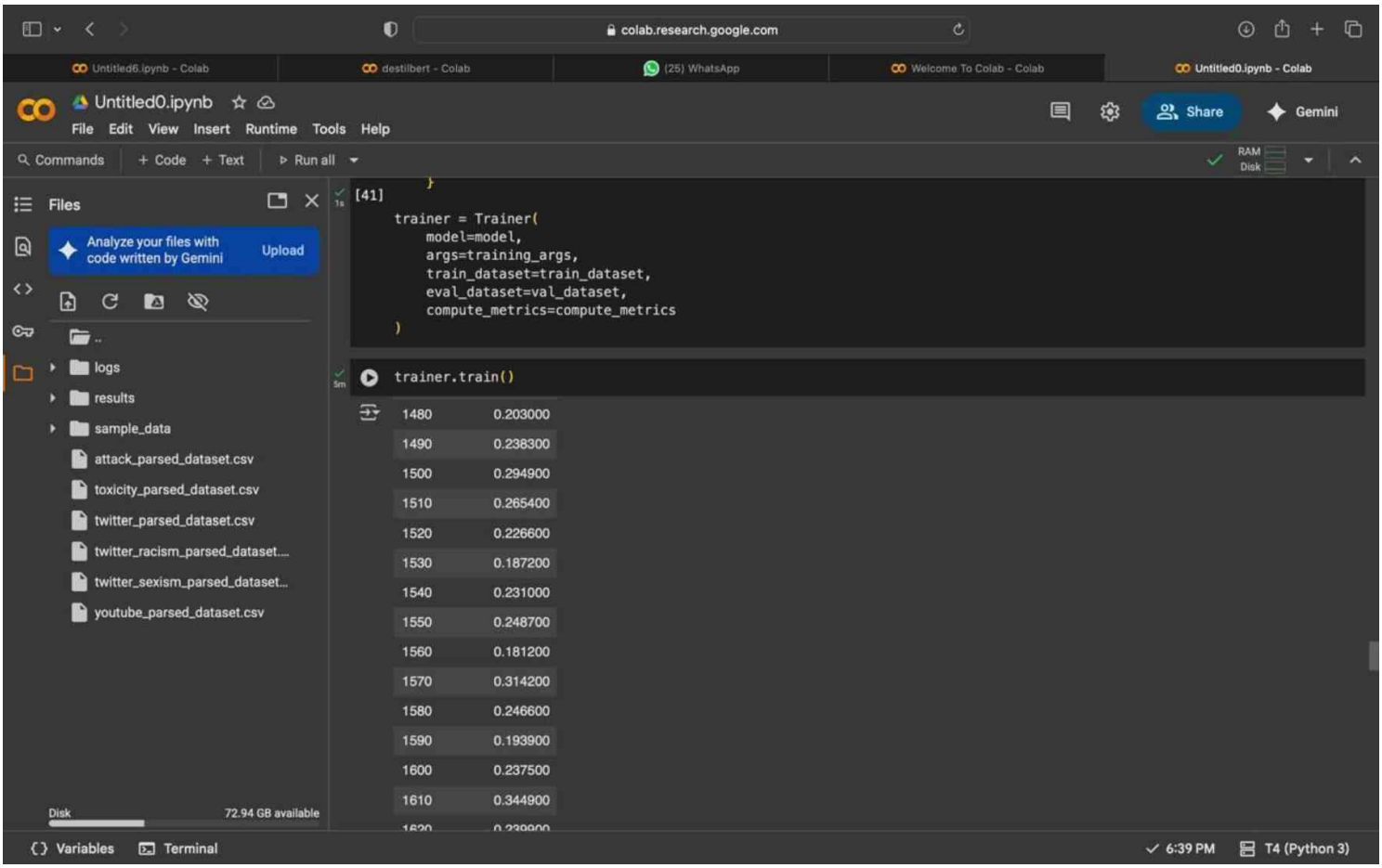

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
# Define all file names and labels
files = {
"toxicity_parsed_dataset.csv": "toxicity",
"twitter_racism_parsed_dataset.csv": "racism",
"twitter_sexism_parsed_dataset.csv": "sexism",
"youtube_parsed_dataset.csv": "youtube"
}
all_dfs = []
for file, label in files.items():
df = pd.read_csv(file)
text_col = None
for col in df.columns:
if 'text' in col.lower():
text_col = col
break
if not text_col:
raise ValueError(f"Couldn't find a text column in {file}")
df = df[[text_col]].dropna().copy()
df.columns = ['text']
df['label'] = label
all_dfs.append(df)
final_df = pd.concat(all_dfs).reset_index(drop=True)
final_df.head()
text label
0 This: :One can make an analogy in mathematical... toxicity
1 ` :Clarification for you (and Zundark's righ... toxicity
2 Elected or Electoral? JHK toxicity
3 `This is such a fun entry. Devotchka I once... toxicity
Variables
Terminal
6:39 PM
T4 (Python 3)

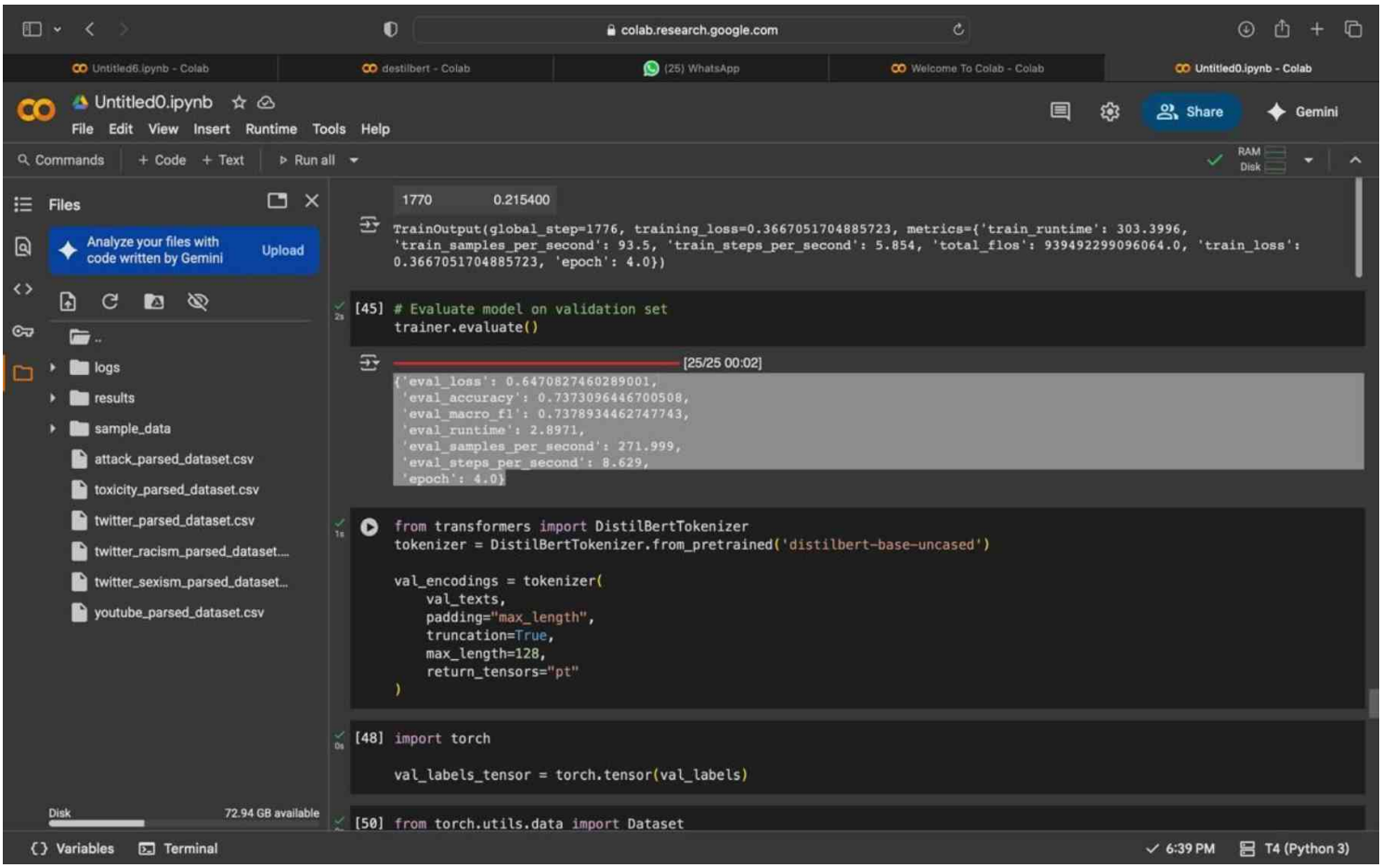

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk 72.94 GB available
from sklearn.utils import resample
import pandas as pd
# Get the smallest class count
min_count = 1970
# Balanced list
balanced = []
# Loop through each label and resample
for label in final_df["label"].unique():
label_df = final_df[final_df["label"] == label]
if len(label_df) > min_count:
resampled = resample(label_df, replace=False, n_samples=min_count, random_state=42)
else:
resampled = resample(label_df, replace=True, n_samples=min_count, random_state=42)
balanced.append(resampled)
# Concatenate all balanced samples
balanced_df = pd.concat(balanced).sample(frac=1, random_state=42).reset_index(drop=True)
# Check new distribution
print("New Label Distribution:")
print(balanced_df["label"].value_counts())
New Label Distribution:
label
toxicity 1970
racism 1970
youtube 1970
sexism 1970
Name: count, dtype: int64
[27] # New label encoding
Variables
Terminal
6:39 PM
T4 (Python 3)

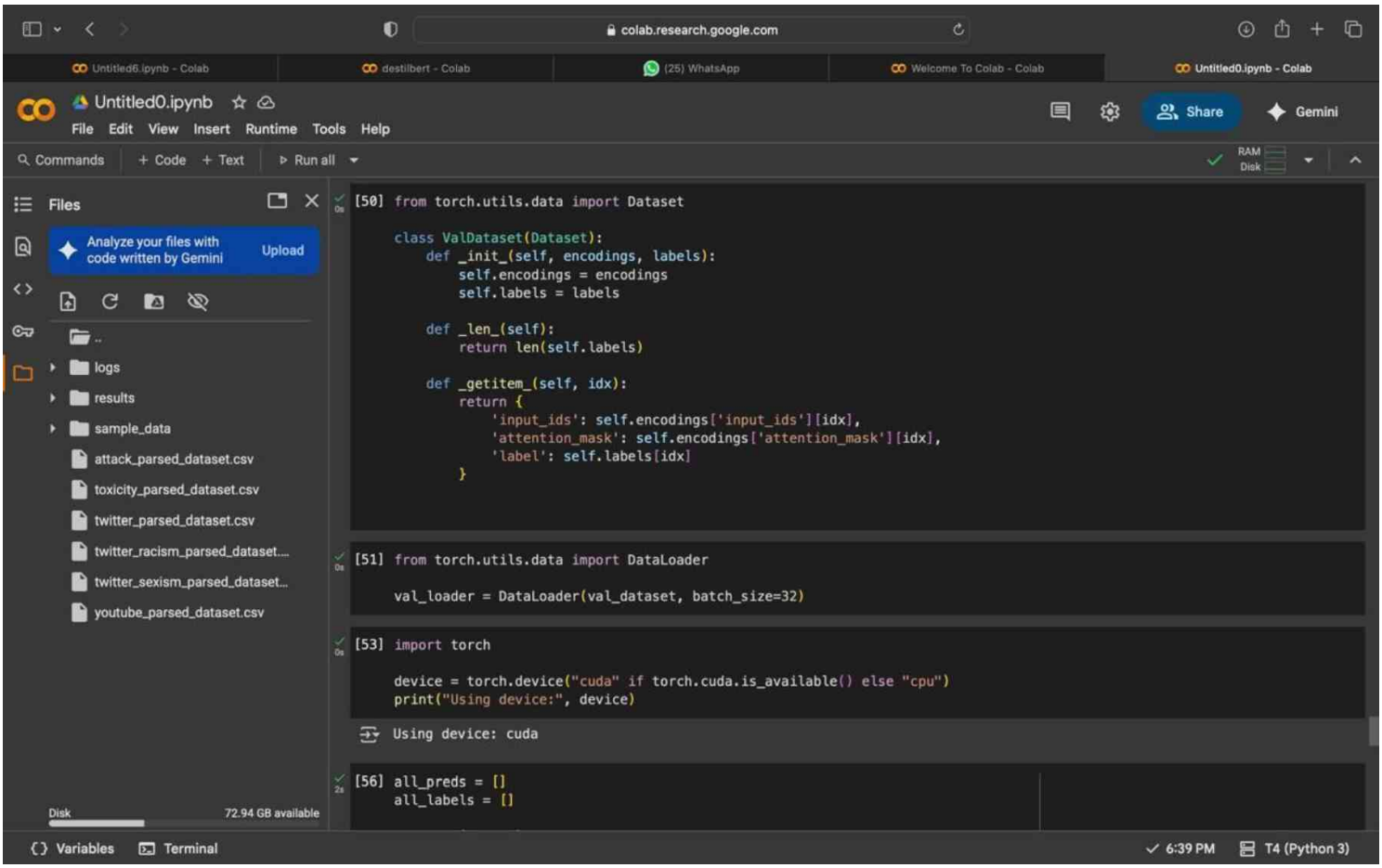

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
[27] # New label encoding
label_map = {'toxicity': 0, 'racism': 1, 'youtube': 2, 'sexism': 3}
balanced_df['label_encoded'] = balanced_df['label'].map(label_map)
import re
def clean(text):
text = re.sub(r"http\S+", "", text)
text = re.sub(r"[^A-Za-z\s]", "", text)
text = re.sub(r"\s+", " ", text).strip()
return text.lower()
balanced_df["clean_text"] = balanced_df["text"].apply(clean)
[33] # Split into train and validation sets (90/10 split)
from sklearn.model_selection import train_test_split
train_texts, val_texts, train_labels, val_labels = train_test_split(
balanced_df['text'].tolist(),
balanced_df['label_encoded'].tolist(),
test_size=0.1,
stratify=balanced_df['label_encoded'],
random_state=42
)
# Load tokenizer
tokenizer = DistilBertTokenizerFast.from_pretrained('distilbert-base-uncased')
# Tokenize train and validation sets
train_encodings = tokenizer(train_texts, truncation=True, padding=True, max_length=128)
val_encodings = tokenizer(val_texts, truncation=True, padding=True, max_length=128)
Variables
Terminal
6:39 PM
T4 (Python 3)

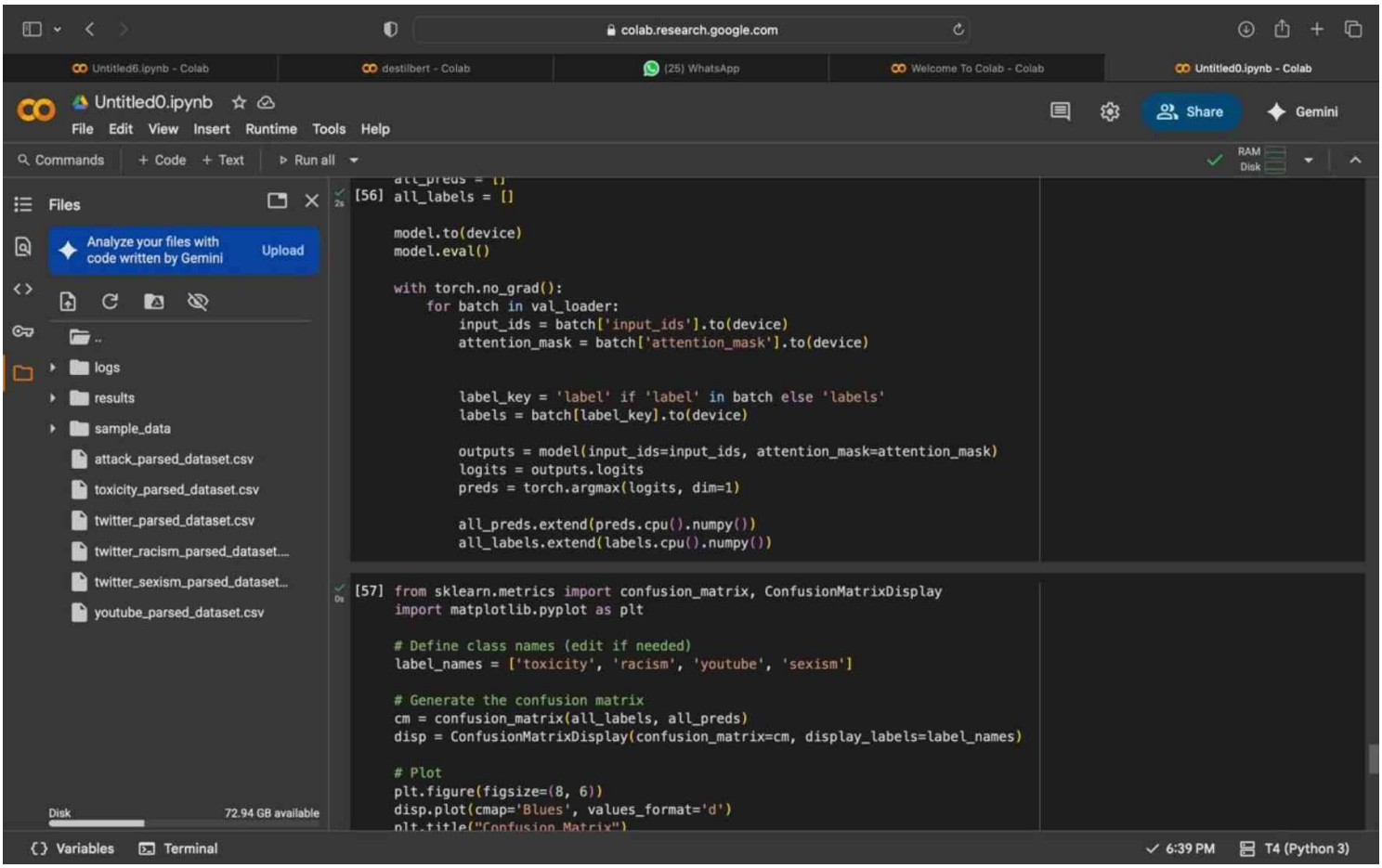

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
[35] !pip install -q datasets
[36] from datasets import Dataset
# Rebuild train and validation into HF-compatible Dataset format
train_dict = {
'text': train_texts,
'label': train_labels
}
val_dict = {
'text': val_texts,
'label': val_labels
}
train_dataset = Dataset.from_dict(train_dict)
val_dataset = Dataset.from_dict(val_dict)
[37] def tokenize_fn(example):
return tokenizer(example['text'], truncation=True, padding='max_length', max_length=128)
# Tokenize datasets
train_dataset = train_dataset.map(tokenize_fn, batched=True)
val_dataset = val_dataset.map(tokenize_fn, batched=True)
# Rename label column for Trainer compatibility
train_dataset = train_dataset.rename_column("label", "labels")
val_dataset = val_dataset.rename_column("label", "labels")
# Set format to PyTorch
train_dataset.set_format(type='torch', columns=['input_ids', 'attention_mask', 'labels'])
val_dataset.set_format(type='torch', columns=['input_ids', 'attention_mask', 'labels'])
Variables
Terminal
6:39 PM
T4 (Python 3)

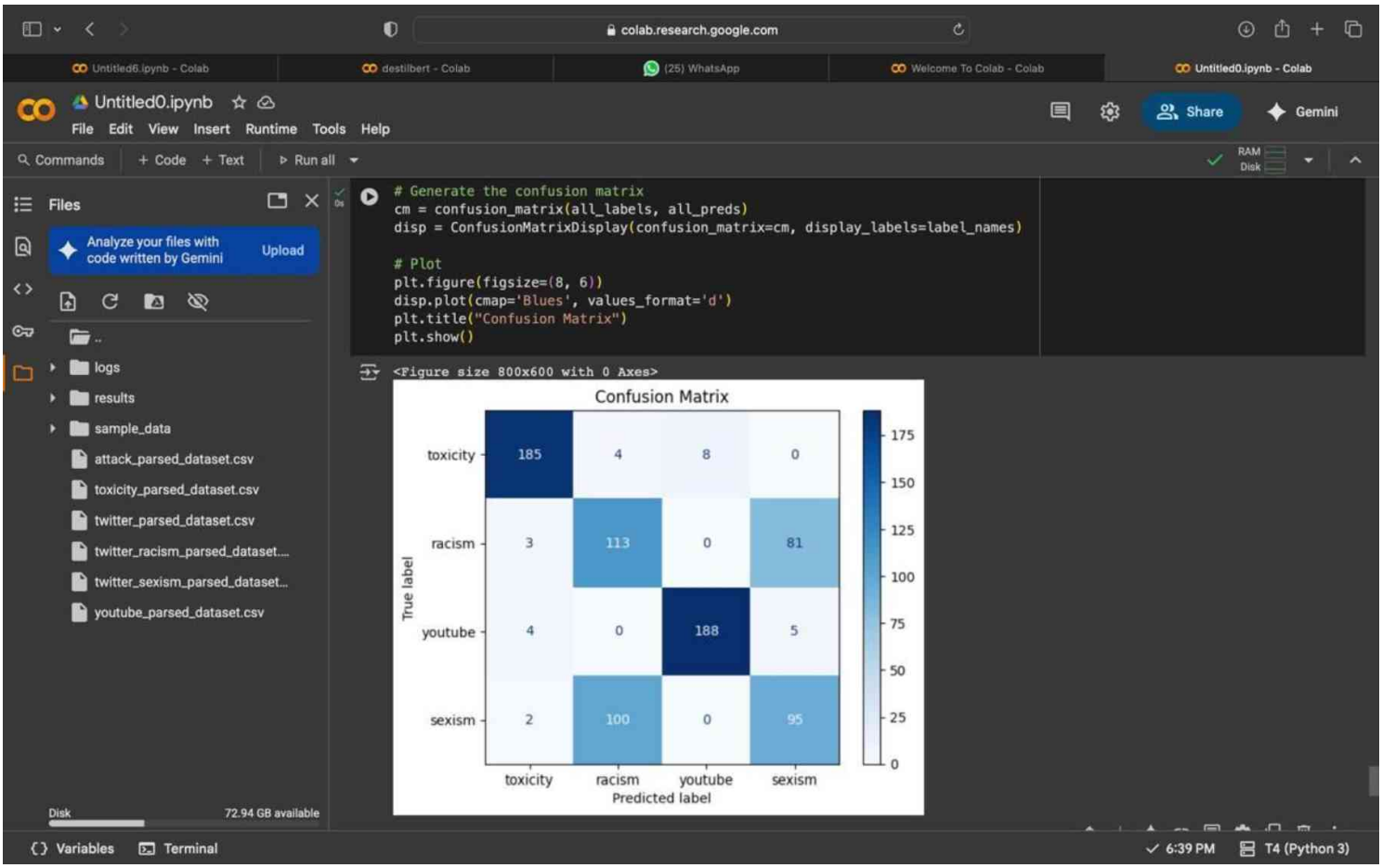

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
Map: 100% 7092/7092 [00:07<00:00, 1001.17 examples/s]
Map: 100% 788/788 [00:00<00:00, 1085.10 examples/s]
[38]
from transformers import DistilBertForSequenceClassification# Number of unique labels (classes)num_labels = len(set(tra
from transformers import TrainingArguments
training_args = TrainingArguments(
output_dir='./results',
num_train_epochs=4,
per_device_train_batch_size=16,
per_device_eval_batch_size=32,
warmup_steps=100,
weight_decay=0.01,
logging_dir='./logs',
logging_steps=10,
save_strategy="no"
)
Using the `WANDB_DISABLED` environment variable is deprecated and will be removed in v5. Use the --report_to flag to con
[41] from sklearn.metrics import accuracy_score, f1_score
def compute_metrics(p):
preds = p.predictions.argmax(axis=1)
labels = p.label_ids
return {
"accuracy": accuracy_score(labels, preds),
"macro_f1": f1_score(labels, preds, average='macro')
}
Variables
Terminal
6:39 PM
T4 (Python 3)

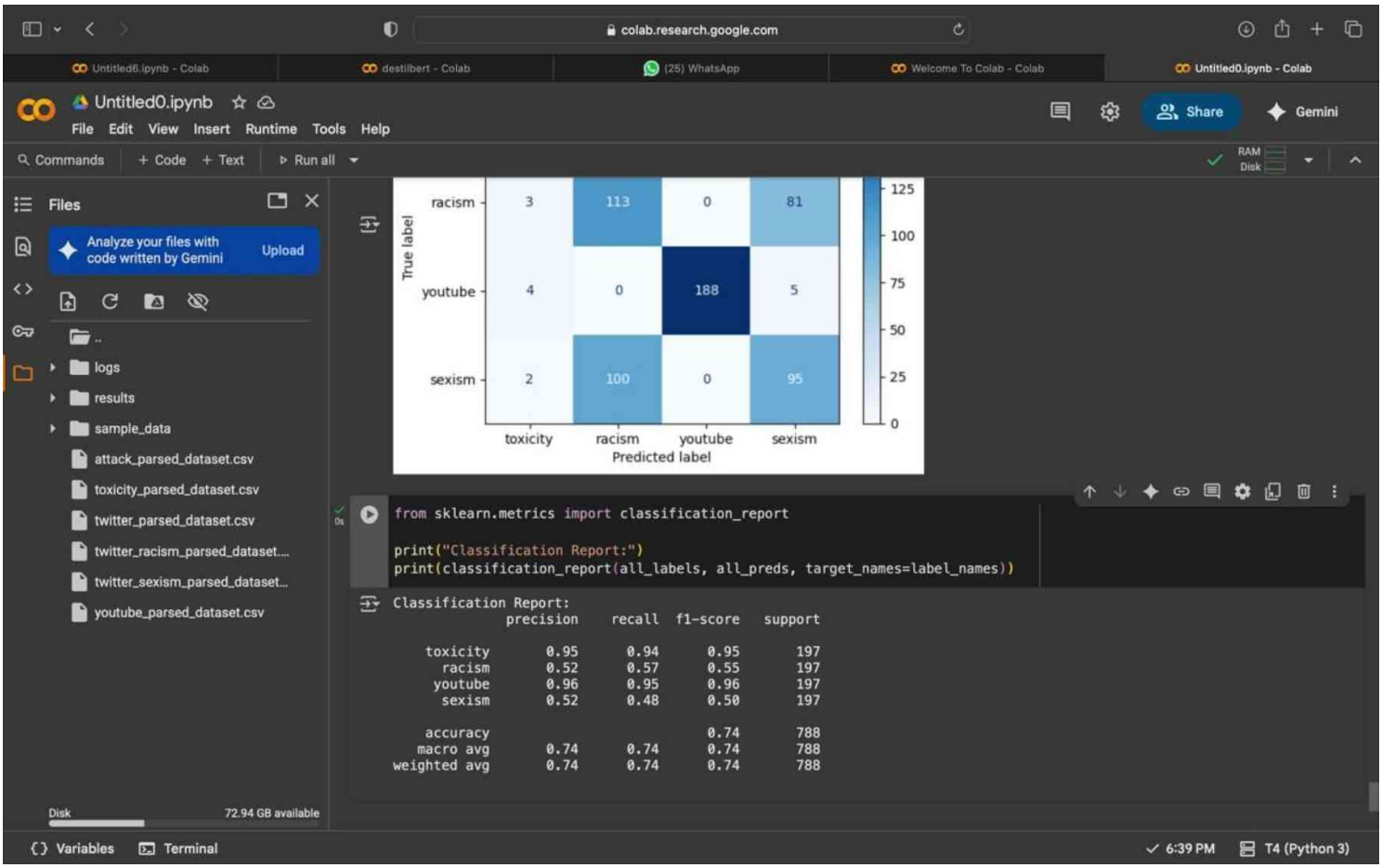

colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk 72.94 GB available
[41]
trainer = Trainer(
model=model,
args=training_args,
train_dataset=train_dataset,
eval_dataset=val_dataset,
compute_metrics=compute_metrics
)
trainer.train()
1480 0.203000
1490 0.238300
1500 0.294900
1510 0.265400
1520 0.226600
1530 0.187200
1540 0.231000
1550 0.248700
1560 0.181200
1570 0.314200
1580 0.246600
1590 0.193900
1600 0.237500
1610 0.344900
Variables Terminal
6:39 PM
T4 (Python 3)


colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk 72.94 GB available
1770 0.215400
TrainOutput(global_step=1776, training_loss=0.3667051704885723, metrics={'train_runtime': 303.3996, 'train_samples_per_second': 93.5, 'train_steps_per_second': 5.854, 'total_flos': 939492299096064.0, 'train_loss': 0.3667051704885723, 'epoch': 4.0})
[45] # Evaluate model on validation set
trainer.evaluate()
[25/25 00:02]
{'eval_loss': 0.6470827460289001,
'eval_accuracy': 0.7373096446700508,
'eval_macro_f1': 0.7378934462747743,
'eval_runtime': 2.8971,
'eval_samples_per_second': 271.999,
'eval_steps_per_second': 8.629,
'epoch': 4.0}
from transformers import DistilBertTokenizer
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')
val_encodings = tokenizer(
val_texts,
padding="max_length",
truncation=True,
max_length=128,
return_tensors="pt"
)
[48] import torch
val_labels_tensor = torch.tensor(val_labels)
[50] from torch.utils.data import Dataset
Variables Terminal
6:39 PM T4 (Python 3)


colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
[50] from torch.utils.data import Dataset
class ValDataset(Dataset):
def _init_(self, encodings, labels):
self.encodings = encodings
self.labels = labels
def _len_(self):
return len(self.labels)
def _getitem_(self, idx):
return {
'input_ids': self.encodings['input_ids'][idx],
'attention_mask': self.encodings['attention_mask'][idx],
'label': self.labels[idx]
}
[51] from torch.utils.data import DataLoader
val_loader = DataLoader(val_dataset, batch_size=32)
[53] import torch
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
print("Using device:", device)
Using device: cuda
[56] all_preds = []
all_labels = []
Variables
Terminal
6:39 PM
T4 (Python 3)


colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
[56] all_labels = []
model.to(device)
model.eval()
with torch.no_grad():
for batch in val_loader:
input_ids = batch['input_ids'].to(device)
attention_mask = batch['attention_mask'].to(device)
label_key = 'label' if 'label' in batch else 'labels'
labels = batch[label_key].to(device)
outputs = model(input_ids=input_ids, attention_mask=attention_mask)
logits = outputs.logits
preds = torch.argmax(logits, dim=1)
all_preds.extend(preds.cpu().numpy())
all_labels.extend(labels.cpu().numpy())
[57] from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt
# Define class names (edit if needed)
label_names = ['toxicity', 'racism', 'youtube', 'sexism']
# Generate the confusion matrix
cm = confusion_matrix(all_labels, all_preds)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=label_names)
# Plot
plt.figure(figsize=(8, 6))
disp.plot(cmap='Blues', values_format='d')
Variables
Terminal
6:39 PM
T4 (Python 3)


colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
# Generate the confusion matrix
cm = confusion_matrix(all_labels, all_preds)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=label_names)
# Plot
plt.figure(figsize=(8, 6))
disp.plot(cmap='Blues', values_format='d')
plt.title("Confusion Matrix")
plt.show()
<Figure size 800x600 with 0 Axes>
Confusion Matrix
toxicity
racism
youtube
sexism
True label
Predicted label
185
4
8
0
3
113
0
81
4
0
188
5
2
100
0
95
175
150
125
100
75
50
25
0
Variables
Terminal
6:39 PM
T4 (Python 3)


colab.research.google.com
Untitled6.ipynb - Colab
destilbert - Colab
(25) WhatsApp
Welcome To Colab - Colab
Untitled0.ipynb - Colab
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Files
Analyze your files with code written by Gemini
Upload
..
logs
results
sample_data
attack_parsed_dataset.csv
toxicity_parsed_dataset.csv
twitter_parsed_dataset.csv
twitter_racism_parsed_dataset....
twitter_sexism_parsed_dataset...
youtube_parsed_dataset.csv
Disk
72.94 GB available
racism
youtube
sexism
True label
3
113
0
81
4
0
188
5
2
100
0
95
toxicity
racism
youtube
sexism
Predicted label
125
100
75
50
25
0
from sklearn.metrics import classification_report
print("Classification Report:")
print(classification_report(all_labels, all_preds, target_names=label_names))
Classification Report:
precision recall f1-score support
toxicity 0.95 0.94 0.95 197
racism 0.52 0.57 0.55 197
youtube 0.96 0.95 0.96 197
sexism 0.52 0.48 0.50 197
accuracy 0.74 788
macro avg 0.74 0.74 0.74 788
weighted avg 0.74 0.74 0.74 788
Variables
Terminal
6:39 PM
T4 (Python 3)