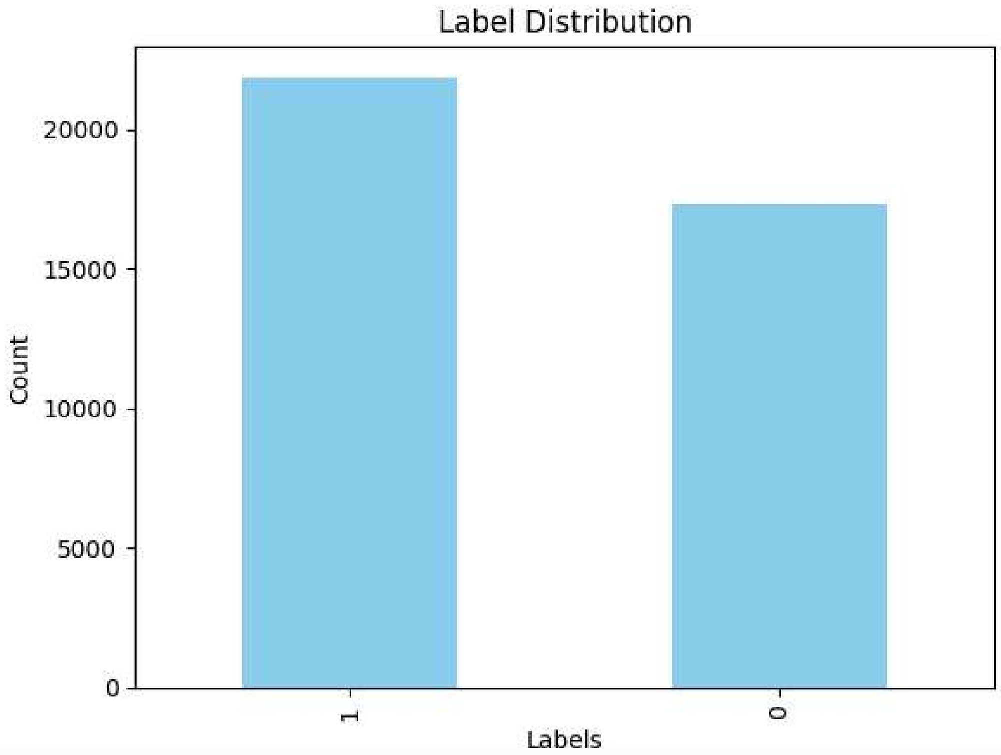
```python
import pandas as pd

df=pd.read_csv('/content/CEAS_08[1].csv')
df.head()
```

| | sender | receiver | date | subject | |
|---|---|---|---|---|---|
| **0** | Young Esposito <Young@iworld.de> | user4@gvc.ceas-challenge.cc | Tue, 05 Aug 2008 16:31:02 -0700 | Never agree to be a loser | |
| **1** | Mok <ipline's1983@icable.ph> | user2.2@gvc.ceas-challenge.cc | Tue, 05 Aug 2008 18:31:03 -0500 | Befriend Jenna Jameson | |
| **2** | Daily Top 10 <Karmandeep-opengevl@universalnet... | user2.9@gvc.ceas-challenge.cc | Tue, 05 Aug 2008 20:28:00 -1200 | CNN.com Daily Top 10 | >+=+ |
| **3** | Michael Parker <ivqrnai@pobox.com> | SpamAssassin Dev <xrh@spamassassin.apache.org> | Tue, 05 Aug 2008 17:31:20 -0600 | Re: svn commit: r619753 - in /spamassassin/tru... | |
| **4** | Gretchen Suggs <externalsep1@loanofficertool.com> | user2.2@gvc.ceas-challenge.cc | Tue, 05 Aug 2008 19:31:21 -0400 | SpecialPricesPharmMoreinfo | |

```python
import pandas as pd
df=pd.read_csv('/content/CEAS_08[1].csv')
df.size

```

```
274078
```

```python
import matplotlib.pyplot as plt

df['label'].value_counts().plot(kind='bar', color='skyblue')
plt.title('Label Distribution')
plt.xlabel('Labels')
plt.ylabel('Count')
plt.show()

```



```
print(df['label'].value_counts())

```

```
label
1    21842
0    17312
Name: count, dtype: int64
```

```
import torch
import torch.nn as nn
from torch.utils.data import Dataset, DataLoader
from transformers import BertTokenizer, BertModel, AdamW
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, f1_score
import re
import numpy as np
import gc

# ----- Data Preprocessing Functions -----
def clean_email_text(text):
    """Clean email text by removing unwanted patterns and normalizing"""
    if pd.isna(text):
        return ""
    text = str(text)
    text = re.sub(r'From:.*\n|To:.*\n|Subject:.*\n|Date:.*\n', '', text)
    text = re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+',
    text = re.sub(r'[^a-zA-Z0-9\s.,!?]', '', text)
    text = re.sub(r'\s+', ' ', text).strip()
    return text.lower()

def preprocess_dataframe(df):
    """Preprocess the entire dataframe"""
    df['combined_text'] = df['subject'].fillna('') + ' ' + df['body'].fillna('')
    df['combined_text'] = df['combined_text'].apply(clean_email_text)
    df['label'] = pd.to_numeric(df['label'], errors='coerce').fillna(0).astype(int)
    return df

# ----- 1. Prepare the Dataset -----
class PhishingEmailDataset(Dataset):
    def __init__(self, texts, labels, tokenizer, max_length=256):
        self.texts = texts
        self.labels = labels
        self.tokenizer = tokenizer
        self.max_length = max_length

    def __len__(self):
        return len(self.texts)

    def __getitem__(self, idx):
        text = str(self.texts[idx])
        label = self.labels[idx]
        if len(text.strip()) == 0:
            text = "[EMPTY_EMAIL]"
        encoding = self.tokenizer.encode_plus(
            text,
            add_special_tokens=True,
            max_length=self.max_length,
            truncation=True,
            padding='max_length',
            return_token_type_ids=False,
            return_attention_mask=True,
            return_tensors='pt'
        )
        return {
            'input_ids': encoding['input_ids'].flatten(),
            'attention_mask': encoding['attention_mask'].flatten(),
            'label': torch.tensor(label, dtype=torch.long)
        }

# ----- 2. Define the Hybrid BERT-GRU Model -----
class BertGRUClassifier(nn.Module):
    def __init__(self, n_classes, dropout_rate=0.3, gru_hidden_size=128, num_gru_layers=1):
        super(BertGRUClassifier, self).__init__()
        self.bert = BertModel.from_pretrained('bert-base-uncased')
        self.gru = nn.GRU(input_size=self.bert.config.hidden_size,
                          hidden_size=gru_hidden_size,
                          num_layers=num_gru_layers,
                          batch_first=True,
                          bidirectional=True)
        self.dropout = nn.Dropout(dropout_rate)
        self.out = nn.Linear(gru_hidden_size * 2, n_classes)

    def forward(self, input_ids, attention_mask):
        bert_outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask)
        sequence_output = bert_outputs.last_hidden_state
        gru_output, _ = self.gru(sequence_output)
        pooled_output = torch.mean(gru_output, dim=1)
        pooled_output = self.dropout(pooled_output)
        logits = self.out(pooled_output)
        return logits

```

```
# ----- 3. Load and Preprocess Dataset -----
df = pd.read_csv("/content/CEAS_08[1].csv", escapechar='\\')
print("Original data shape:", df.shape)

df_processed = preprocess_dataframe(df)
print("Processed data shape:", df_processed.shape)
print("Label distribution:", df_processed['label'].value_counts())

train_texts, val_texts, train_labels, val_labels = train_test_split(
    df_processed['combined_text'],
    df_processed['label'],
    test_size=0.2,
    random_state=42,
    stratify=df_processed['label']
)

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
train_dataset = PhishingEmailDataset(train_texts.tolist(), train_labels.tolist(), tokenizer)
val_dataset = PhishingEmailDataset(val_texts.tolist(), val_labels.tolist(), tokenizer)

batch_size = 8  # Reduced batch size to prevent memory overload
train_loader = DataLoader(train_dataset, batch_size=batch_size, shuffle=True, num_workers=2)
val_loader = DataLoader(val_dataset, batch_size=batch_size, num_workers=2)

# ----- 4. Training Setup -----
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
model = BertGRUClassifier(n_classes=2).to(device)
optimizer = AdamW(model.parameters(), lr=2e-5)
criterion = nn.CrossEntropyLoss()

def train_epoch(model, data_loader, optimizer, criterion, device):
    model.train()
    total_loss = 0
    for i, batch in enumerate(data_loader):
        input_ids = batch['input_ids'].to(device)
        attention_mask = batch['attention_mask'].to(device)
        labels = batch['label'].to(device)

        optimizer.zero_grad()
        outputs = model(input_ids=input_ids, attention_mask=attention_mask)
        loss = criterion(outputs, labels)

        # Gradient clipping to prevent exploding gradients
        torch.nn.utils.clip_grad_norm_(model.parameters(), max_norm=1.0)

        loss.backward()
        optimizer.step()

        total_loss += loss.item()

        # Clear memory
        del input_ids, attention_mask, labels, outputs, loss
        torch.cuda.empty_cache()

        if i % 10 == 0:  # Print progress every 10 batches
            print(f"Batch {i}/{len(data_loader)} processed")

    return total_loss / len(data_loader)

```

```
def eval_model(model, data_loader, device):
    model.eval()
    predictions = []
    true_labels = []
    with torch.no_grad():
        for batch in data_loader:
            input_ids = batch['input_ids'].to(device)
            attention_mask = batch['attention_mask'].to(device)
            labels = batch['label'].to(device)
            outputs = model(input_ids=input_ids, attention_mask=attention_mask)
            _, preds = torch.max(outputs, dim=1)
            predictions.extend(preds.cpu().numpy())
            true_labels.extend(labels.cpu().numpy())

            # Clear memory
            del input_ids, attention_mask, labels, outputs
            torch.cuda.empty_cache()

    return accuracy_score(true_labels, predictions), f1_score(true_labels, predictions, average='weigh

# ----- 5. Training Loop -----
epochs = 10
for epoch in range(epochs):
    print(f"Starting Epoch {epoch+1}/{epochs}")
    train_loss = train_epoch(model, train_loader, optimizer, criterion, device)
    val_accuracy, val_f1 = eval_model(model, val_loader, device)
    print(f"Epoch {epoch+1}/{epochs} | Train Loss: {train_loss:.4f} | Val Acc: {val_accuracy:.4f} | Va

    # Garbage collection
    gc.collect()
    torch.cuda.empty_cache()

print("Training completed!")
```

```
Processed data shape: (39154, 8)
Label distribution: label
1    21842
0    17312
Name: count, dtype: int64
tokenizer_config.json: 100%                     48.0/48.0 [00:00<00:00, 3.81kB/s]
vocab.txt: 100%                                 232k/232k [00:00<00:00, 2.86MB/s]
tokenizer.json: 100%                            466k/466k [00:00<00:00, 4.41MB/s]
config.json: 100%                               570/570 [00:00<00:00, 45.0kB/s]
model.safetensors: 100%                         440M/440M [00:02<00:00, 228MB/s]
Starting Epoch 1/10
/usr/local/lib/python3.11/dist-packages/transformers/optimization.py:640: FutureWarning: This implem
  warnings.warn(
Batch 0/3916 processed
Batch 10/3916 processed
Batch 20/3916 processed
Batch 30/3916 processed
Batch 40/3916 processed
Batch 50/3916 processed
Batch 60/3916 processed
Batch 70/3916 processed
Batch 80/3916 processed
Batch 90/3916 processed
Batch 100/3916 processed
Batch 110/3916 processed
Batch 120/3916 processed
Batch 130/3916 processed
Batch 140/3916 processed
Batch 150/3916 processed
Batch 160/3916 processed
Batch 170/3916 processed
Batch 180/3916 processed
Batch 190/3916 processed
Batch 200/3916 processed
Batch 210/3916 processed
Batch 220/3916 processed
Batch 230/3916 processed
Batch 240/3916 processed
Batch 250/3916 processed
Batch 260/3916 processed
Batch 270/3916 processed
Batch 280/3916 processed
Batch 290/3916 processed
Batch 300/3916 processed
Batch 310/3916 processed
Batch 320/3916 processed
Batch 330/3916 processed
Batch 340/3916 processed
Batch 350/3916 processed
Batch 360/3916 processed
Batch 370/3916 processed
Batch 380/3916 processed
Batch 390/3916 processed
Batch 400/3916 processed
Batch 410/3916 processed
Batch 420/3916 processed
Batch 430/3916 processed
Batch 440/3916 processed
Batch 450/3916 processed
Batch 460/3916 processed
Batch 470/3916 processed
Batch 480/3916 processed
Batch 490/3916 processed
Batch 500/3916 processed
Batch 510/3916 processed
Batch 520/3916 processed
Batch 530/3916 processed
Batch 540/3916 processed
Batch 550/3916 processed
Batch 560/3916 processed
Batch 570/3916 processed
Batch 580/3916 processed
Batch 590/3916 processed
Batch 600/3916 processed
Batch 610/3916 processed
Batch 620/3916 processed
Batch 630/3916 processed
Batch 640/3916 processed
Batch 650/3916 processed
Batch 660/3916 processed
Batch 670/3916 processed
Batch 680/3916 processed
Batch 690/3916 processed
Batch 700/3916 processed
Batch 710/3916 processed
Batch 720/3916 processed
Batch 730/3916 processed
Batch 740/3916 processed
Batch 750/3916 processed
Batch 760/3916 processed
Batch 770/3916 processed
Batch 780/3916 processed
Batch 790/3916 processed
Batch 800/3916 processed
Batch 810/3916 processed
Batch 820/3916 processed
Batch 830/3916 processed
Batch 840/3916 processed
Batch 850/3916 processed
Batch 860/3916 processed
Batch 870/3916 processed
Batch 880/3916 processed
Batch 890/3916 processed
Batch 900/3916 processed
Batch 910/3916 processed
Batch 920/3916 processed
Batch 930/3916 processed
Batch 940/3916 processed
Batch 950/3916 processed
Batch 960/3916 processed
Batch 970/3916 processed
Batch 980/3916 processed
Batch 990/3916 processed
Batch 1000/3916 processed
Batch 1010/3916 processed
Batch 1020/3916 processed
Batch 1030/3916 processed
Batch 1040/3916 processed
Batch 1050/3916 processed
Batch 1060/3916 processed
Batch 1070/3916 processed
Batch 1080/3916 processed
Batch 1090/3916 processed
Batch 1100/3916 processed
Batch 1110/3916 processed
Batch 1120/3916 processed
Batch 1130/3916 processed
Batch 1140/3916 processed
Batch 1150/3916 processed
Batch 1160/3916 processed
Batch 1170/3916 processed
Batch 1180/3916 processed
Batch 1190/3916 processed
Batch 1200/3916 processed
Batch 1210/3916 processed
Batch 1220/3916 processed
Batch 1230/3916 processed
Batch 1240/3916 processed
Batch 1250/3916 processed
Batch 1260/3916 processed
Batch 1270/3916 processed
Batch 1280/3916 processed
Batch 1290/3916 processed
Batch 1300/3916 processed
Batch 1310/3916 processed
Batch 1320/3916 processed
Batch 1330/3916 processed
Batch 1340/3916 processed
Batch 1350/3916 processed
Batch 1360/3916 processed
Batch 1370/3916 processed
Batch 1380/3916 processed
Batch 1390/3916 processed
Batch 1400/3916 processed
Batch 1410/3916 processed
Batch 1420/3916 processed
Batch 1430/3916 processed
Batch 1440/3916 processed
Batch 1450/3916 processed
Batch 1460/3916 processed
Batch 1470/3916 processed
Batch 1480/3916 processed
Batch 1490/3916 processed
Batch 1500/3916 processed
Batch 1510/3916 processed
Batch 1520/3916 processed
Batch 1530/3916 processed
Batch 1540/3916 processed
Batch 1550/3916 processed
Batch 1560/3916 processed
Batch 1570/3916 processed
Batch 1580/3916 processed
Batch 1590/3916 processed
Batch 1600/3916 processed
Batch 1610/3916 processed
Batch 1620/3916 processed
Batch 1630/3916 processed
Batch 1640/3916 processed
Batch 1650/3916 processed
Batch 1660/3916 processed
Batch 1670/3916 processed
Batch 1680/3916 processed
Batch 1690/3916 processed
Batch 1700/3916 processed
Batch 1710/3916 processed
Batch 1720/3916 processed
Batch 1730/3916 processed
Batch 1740/3916 processed
Batch 1750/3916 processed
Batch 1760/3916 processed
Batch 1770/3916 processed
Batch 1770/3916 processed
Batch 1780/3916 processed
Batch 1790/3916 processed
Batch 1800/3916 processed
Batch 1810/3916 processed
Batch 1820/3916 processed
Batch 1830/3916 processed
Batch 1840/3916 processed
Batch 1850/3916 processed
Batch 1860/3916 processed
Batch 1870/3916 processed
Batch 1880/3916 processed
Batch 1890/3916 processed
Batch 1900/3916 processed
Batch 1910/3916 processed
Batch 1920/3916 processed
Batch 1930/3916 processed
Batch 1940/3916 processed
Batch 1950/3916 processed
Batch 1960/3916 processed
Batch 1970/3916 processed
Batch 1980/3916 processed
Batch 1990/3916 processed
Batch 2000/3916 processed
Batch 2010/3916 processed
Batch 2020/3916 processed
Batch 2030/3916 processed
Batch 2040/3916 processed
Batch 2050/3916 processed
Batch 2060/3916 processed
Batch 2070/3916 processed
Batch 2080/3916 processed
Batch 2090/3916 processed
Batch 2100/3916 processed
Batch 2110/3916 processed
Batch 2120/3916 processed
Batch 2130/3916 processed
Batch 2140/3916 processed
Batch 2150/3916 processed
Batch 2160/3916 processed
Batch 2170/3916 processed
Batch 2180/3916 processed
Batch 2190/3916 processed
Batch 2200/3916 processed
Batch 2210/3916 processed
Batch 2220/3916 processed
Batch 2230/3916 processed
Batch 2240/3916 processed
Batch 2250/3916 processed
Batch 2260/3916 processed
Batch 2270/3916 processed
Batch 2280/3916 processed
Batch 2290/3916 processed
Batch 2300/3916 processed
Batch 2310/3916 processed
Batch 2320/3916 processed
Batch 2330/3916 processed
Batch 2340/3916 processed
Batch 2350/3916 processed
Batch 2360/3916 processed
Batch 2370/3916 processed
Batch 2380/3916 processed
Batch 2390/3916 processed
Batch 2400/3916 processed
Batch 2410/3916 processed
Batch 2420/3916 processed
Batch 2430/3916 processed
Batch 2440/3916 processed
Batch 2450/3916 processed
Batch 2460/3916 processed
Batch 2470/3916 processed
Batch 2480/3916 processed
Batch 2490/3916 processed
Batch 2500/3916 processed
Batch 2510/3916 processed
Batch 2520/3916 processed
Batch 2530/3916 processed
Batch 2540/3916 processed
Batch 2550/3916 processed
Batch 2560/3916 processed
Batch 2570/3916 processed
Batch 2580/3916 processed
Batch 2590/3916 processed
Batch 2600/3916 processed
Batch 2610/3916 processed
Batch 2620/3916 processed
Batch 2630/3916 processed
Batch 2640/3916 processed
Batch 2650/3916 processed
Batch 2660/3916 processed
Batch 2670/3916 processed
Batch 2680/3916 processed
Batch 2690/3916 processed
Batch 2700/3916 processed
Batch 2710/3916 processed
Batch 2720/3916 processed
Batch 2730/3916 processed
Batch 2740/3916 processed
Batch 2750/3916 processed
Batch 2760/3916 processed
Batch 2770/3916 processed
Batch 2780/3916 processed
Batch 2790/3916 processed
Batch 2800/3916 processed
Batch 2810/3916 processed
Batch 2820/3916 processed
Batch 2830/3916 processed
Batch 2840/3916 processed
Batch 2850/3916 processed
Batch 2860/3916 processed
Batch 2870/3916 processed
Batch 2880/3916 processed
Batch 2890/3916 processed
Batch 2900/3916 processed
Batch 2910/3916 processed
Batch 2920/3916 processed
Batch 2930/3916 processed
Batch 2940/3916 processed
Batch 2950/3916 processed
Batch 2960/3916 processed
Batch 2970/3916 processed
Batch 2980/3916 processed
Batch 2990/3916 processed
Batch 3000/3916 processed
Batch 3010/3916 processed
Batch 3020/3916 processed
Batch 3030/3916 processed
Batch 3040/3916 processed
Batch 3050/3916 processed
Batch 3060/3916 processed
Batch 3070/3916 processed
Batch 3080/3916 processed
```

```
        Batch 3080/3916 processed
        Batch 3090/3916 processed
        Batch 3100/3916 processed
        Batch 3110/3916 processed
        Batch 3120/3916 processed
        Batch 3130/3916 processed
        Batch 3140/3916 processed
        Batch 3150/3916 processed
        Batch 3160/3916 processed
        Batch 3170/3916 processed
        Batch 3180/3916 processed
        Batch 3190/3916 processed
        Batch 3200/3916 processed
        Batch 3210/3916 processed
        Batch 3220/3916 processed
        Batch 3230/3916 processed
        Batch 3240/3916 processed
        Batch 3250/3916 processed
        Batch 3260/3916 processed
        Batch 3270/3916 processed
        Batch 3280/3916 processed
        Batch 3290/3916 processed
        Batch 3300/3916 processed
        Batch 3310/3916 processed
        Batch 3320/3916 processed
        Batch 3330/3916 processed
        Batch 3340/3916 processed
        Batch 3350/3916 processed
        Batch 3360/3916 processed
        Batch 3370/3916 processed
        Batch 3380/3916 processed
        Batch 3390/3916 processed
        Batch 3400/3916 processed
        Batch 3410/3916 processed
        Batch 3420/3916 processed
        Batch 3430/3916 processed
        Batch 3440/3916 processed
        Batch 3450/3916 processed
        Batch 3460/3916 processed
        Batch 3470/3916 processed
        Batch 3480/3916 processed
        Batch 3490/3916 processed
        Batch 3500/3916 processed
        Batch 3510/3916 processed
        Batch 3520/3916 processed
        Batch 3530/3916 processed
        Batch 3540/3916 processed
        Batch 3550/3916 processed
        Batch 3560/3916 processed
        Batch 3570/3916 processed
        Batch 3580/3916 processed
        Batch 3590/3916 processed
        Batch 3600/3916 processed
        Batch 3610/3916 processed
        Batch 3620/3916 processed
        Batch 3630/3916 processed
        Batch 3640/3916 processed
        Batch 3650/3916 processed
        Batch 3660/3916 processed
        Batch 3670/3916 processed
        Batch 3680/3916 processed
        Batch 3690/3916 processed
        Batch 3700/3916 processed
        Batch 3710/3916 processed
        Batch 3720/3916 processed
        Batch 3730/3916 processed
```

```
Batch 3740/3916 processed
Batch 3750/3916 processed
Batch 3760/3916 processed
Batch 3770/3916 processed
Batch 3780/3916 processed
Batch 3790/3916 processed
Batch 3800/3916 processed
Batch 3810/3916 processed
Batch 3820/3916 processed
Batch 3830/3916 processed
Batch 3840/3916 processed
Batch 3850/3916 processed
Batch 3860/3916 processed
Batch 3870/3916 processed
Batch 3880/3916 processed
Batch 3890/3916 processed
Batch 3900/3916 processed
Batch 3910/3916 processed
Epoch 1/10 | Train Loss: 0.0288 | Val Acc: 0.9980 | Val F1: 0.9980
Starting Epoch 2/10
Batch 0/3916 processed
Batch 10/3916 processed
Batch 20/3916 processed
Batch 30/3916 processed
Batch 40/3916 processed
Batch 50/3916 processed
Batch 60/3916 processed
Batch 70/3916 processed
Batch 80/3916 processed
Batch 90/3916 processed
Batch 100/3916 processed
Batch 110/3916 processed
Batch 120/3916 processed
Batch 130/3916 processed
Batch 140/3916 processed
Batch 150/3916 processed
Batch 160/3916 processed
Batch 170/3916 processed
Batch 180/3916 processed
Batch 190/3916 processed
Batch 200/3916 processed
Batch 210/3916 processed
Batch 220/3916 processed
Batch 230/3916 processed
Batch 240/3916 processed
Batch 250/3916 processed
Batch 260/3916 processed
Batch 270/3916 processed
Batch 280/3916 processed
Batch 290/3916 processed
Batch 300/3916 processed
Batch 310/3916 processed
Batch 320/3916 processed
Batch 330/3916 processed
Batch 340/3916 processed
Batch 350/3916 processed
Batch 360/3916 processed
Batch 370/3916 processed
Batch 380/3916 processed
Batch 390/3916 processed
Batch 400/3916 processed
Batch 410/3916 processed
Batch 420/3916 processed
Batch 430/3916 processed
Batch 440/3916 processed
Batch 450/3916 processed
```