

```

# Import necessary libraries
import pandas as pd
import numpy as np
import nltk
import string
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report, confusion_matrix, ConfusionMa
import joblib
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Download NLTK resources
nltk.download('stopwords')

# Load datasets
train_data = pd.read_csv('/content/train.csv')
test_data = pd.read_csv('/content/test.csv')

# Preview datasets
print("Training Data Preview:")
print(train_data.head(), "\n")
print("Test Data Preview:")
print(test_data.head(), "\n")

# Preprocessing function for cleaning and tokenizing text
def preprocess_text(text):
    if isinstance(text, str): # Ensure input is string
        # Remove punctuation and digits
        text = ''.join([char for char in text if char not in string.punctuation and not char.isdigit()])
        # Convert to lowercase
        text = text.lower()
        # Remove stopwords
        stop_words = set(nltk.corpus.stopwords.words('english'))
        text = ' '.join([word for word in text.split() if word not in stop_words])
        # Apply stemming
        stemmer = nltk.stem.PorterStemmer()
        text = ' '.join([stemmer.stem(word) for word in text.split()])
        return text
    else:
        return ""

# Apply preprocessing
train_data['crimeadditionalinfo'] = train_data['crimeadditionalinfo'].apply(preprocess_text)
test_data['crimeadditionalinfo'] = test_data['crimeadditionalinfo'].apply(preprocess_text)

# Feature extraction using TF-IDF
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_train = tfidf_vectorizer.fit_transform(train_data['crimeadditionalinfo'])
X_test = tfidf_vectorizer.transform(test_data['crimeadditionalinfo'])

# Labels
y_train = train_data['category']
y_test = test_data['category']

# Train-test split
X_train_split, X_val_split, y_train_split, y_val_split = train_test_split(X_train, y_train, test_size=0.2, random_state=42)

# Train model using SVM
model = SVC(kernel='linear') # You can change the kernel as needed
model.fit(X_train_split, y_train_split)

# Evaluate on validation set
y_val_pred = model.predict(X_val_split)
print("\nValidation Set Performance:")
print(classification_report(y_val_split, y_val_pred))

# Evaluate on test set
y_test_pred = model.predict(X_test)
test_metrics = classification_report(y_test, y_test_pred, output_dict=True)
test_accuracy = test_metrics["accuracy"]
print("\nTest Set Performance:")
print(f"Accuracy: {test_accuracy:.4f}")
print(classification_report(y_test, y_test_pred))

# Save the model and vectorizer
joblib.dump(model, 'text_classification_model_svm.pkl')
joblib.dump(tfidf_vectorizer, 'tfidf_vectorizer.pkl')

```

```
# Confusion Matrix
class_labels = sorted(test_data['category'].unique())
conf_matrix = confusion_matrix(y_test, y_test_pred, labels=class_labels)
plt.figure(figsize=(10, 8))
ConfusionMatrixDisplay(confusion_matrix=conf_matrix, display_labels=class_labels).plot(cmap='viridis', ax=plt.gca())
plt.title('Confusion Matrix')
plt.show()

# Word Cloud
all_text = ' '.join(train_data['crimeadditionalinfo'].astype(str))
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Crime Descriptions')
plt.show()

# Bar Plot for Category Distribution
category_counts = train_data['category'].value_counts()
plt.figure(figsize=(10, 6))
category_counts.plot(kind='bar', color='skyblue')
plt.title('Distribution of Crime Categories')
plt.xlabel('Category')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.show()

# Save metrics
metrics_df = pd.DataFrame([{'accuracy': test_metrics['accuracy'],
                             'precision': test_metrics['weighted avg']['precision'],
                             'recall': test_metrics['weighted avg']['recall'],
                             'f1_score': test_metrics['weighted avg']['f1-score']}])
metrics_df.to_csv('evaluation_metrics_svm.csv', index=False)
print("\nMetrics saved to evaluation_metrics_svm.csv.")
```

[nltk_data] Downloading package stopwords to /root/nltk_data...
 [nltk_data] Unzipping corpora/stopwords.zip.
 Training Data Preview:

	category	sub_category \
0	Online and Social Media Related Crime	Cyber Bullying Stalking Sexting
1	Online Financial Fraud	Fraud CallVishing
2	Online Gambling Betting	Online Gambling Betting
3	Online and Social Media Related Crime	Online Job Fraud
4	Online Financial Fraud	Fraud CallVishing

crimeadditionalinfo

```

0 I had continue received random calls and abusi...
1 The above fraudster is continuously messaging ...
2 He is acting like a police and demanding for m...
3 In apna Job I have applied for job interview f...
4 I received a call from lady stating that she w...
```

Test Data Preview:

	category \
0	RapeGang Rape RGRSexually Abusive Content
1	Online Financial Fraud
2	Cyber Attack/ Dependent Crimes
3	Online Financial Fraud
4	Any Other Cyber Crime

	sub_category \
0	NaN
1	DebitCredit Card FraudSim Swap Fraud
2	SQL Injection
3	Fraud CallVishing
4	Other

crimeadditionalinfo

```

0 Sir namaskar mein Ranjit Kumar PatraPaise neh...
1 KOTAK MAHINDRA BANK FRAUD\r\nFRAUD AMOUNT
2 The issue actually started when I got this ema...
3 I am amit kumar from karwi chitrakoot I am tot...
4 I have ordered saree and blouse from rinki s...
```

Validation Set Performance:

```

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

	precision	recall	f1-score	support
Any Other Cyber Crime	0.45	0.21	0.28	2091
Child Pornography CPChild Sexual Abuse Material CSAM	0.75	0.30	0.43	69
Cryptocurrency Crime	0.68	0.40	0.50	96
Cyber Attack/ Dependent Crimes	1.00	1.00	1.00	765
Cyber Terrorism	0.00	0.00	0.00	31
Hacking Damage to computercomputer system etc	0.46	0.23	0.30	341
Online Cyber Trafficking	0.00	0.00	0.00	34
Online Financial Fraud	0.81	0.96	0.88	11471
Online Gambling Betting	0.80	0.08	0.15	97
Online and Social Media Related Crime	0.56	0.59	0.58	2453
Ransomware	0.33	0.09	0.14	11
RapeGang Rape RGRSexually Abusive Content	1.00	0.91	0.95	565
Report Unlawful Content	0.00	0.00	0.00	1
Sexually Explicit Act	0.57	0.01	0.02	322
Sexually Obscene material	0.53	0.02	0.04	391
accuracy			0.77	18738
macro avg	0.53	0.32	0.35	18738
weighted avg	0.73	0.77	0.73	18738

```

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

Test Set Performance:

Accuracy: 0.7559

```

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined ar
_warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

	precision	recall	f1-score	support
Any Other Cyber Crime	0.45	0.19	0.27	3670
Child Pornography CPChild Sexual Abuse Material CSAM	0.76	0.30	0.43	123
Crime Against Women & Children	0.00	0.00	0.00	4
Cryptocurrency Crime	0.68	0.52	0.59	166

	Cyber Attack/ Dependent Crimes	1.00	1.00	1.00	1261
	Cyber Terrorism	0.00	0.00	0.00	52
Hacking	Damage to computercomputer system etc	0.42	0.22	0.29	592
	Online Cyber Trafficking	0.00	0.00	0.00	61
	Online Financial Fraud	0.80	0.96	0.87	18896
	Online Gambling Betting	0.90	0.07	0.12	134
	Online and Social Media Related Crime	0.56	0.58	0.57	4139
	Ransomware	0.40	0.11	0.17	18
RapeGang Rape RGR	Sexually Abusive Content	1.00	0.91	0.95	912
	Sexually Explicit Act	0.00	0.00	0.00	535
	Sexually Obscene material	0.60	0.04	0.07	666
	accuracy			0.76	31229
	macro avg	0.50	0.33	0.36	31229
	weighted avg	0.71	0.76	0.71	31229



