Cyber Crime Category and Sub Category Classification

Dataset Provided

Training: 93685 rows

Testing: 31230 rows

Training Dataset after removing null values of column 'crimeaditionalinfo': 93665 rows

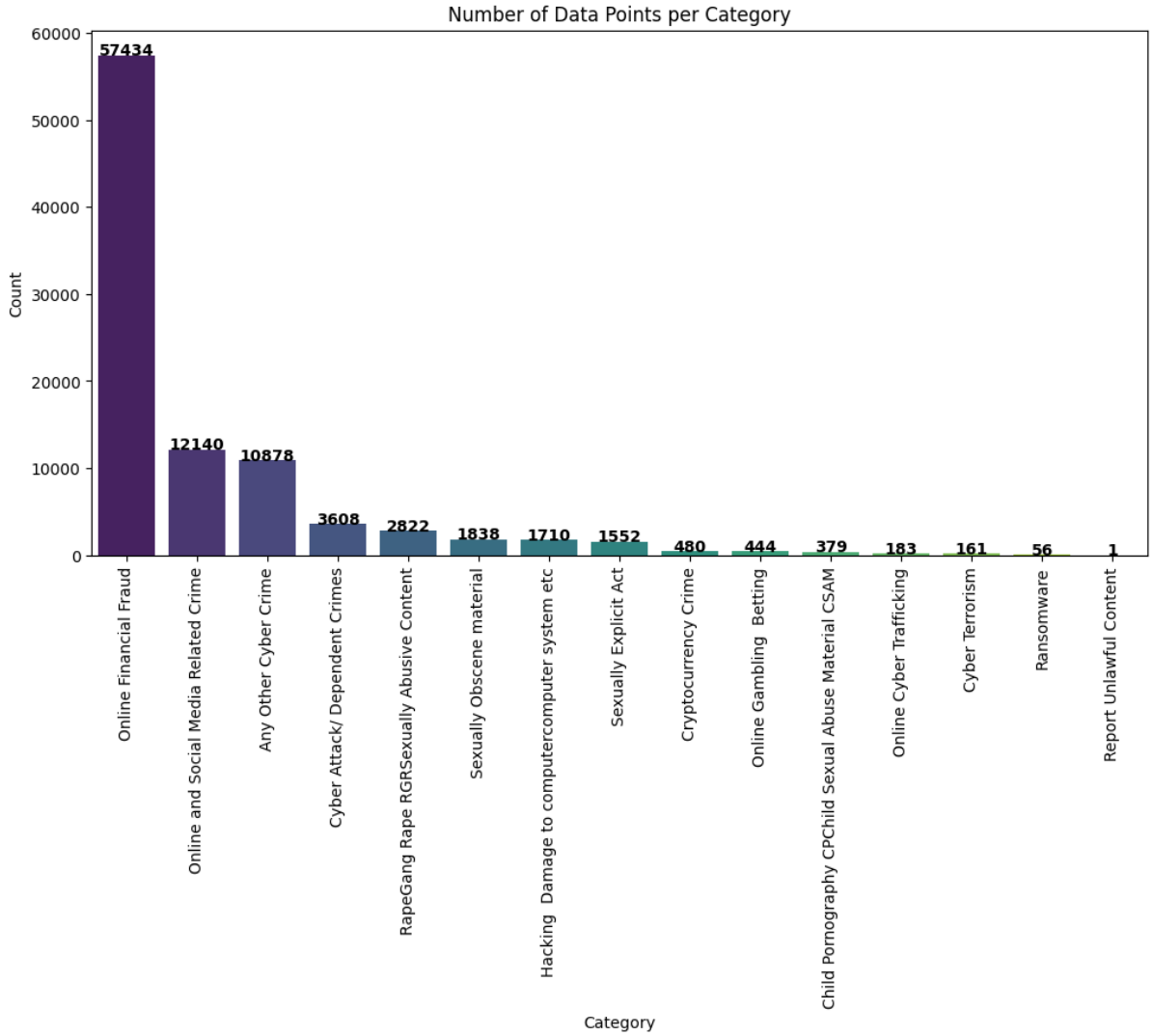
Number of distinct Categories: 15

Number of distinct Categories: 35

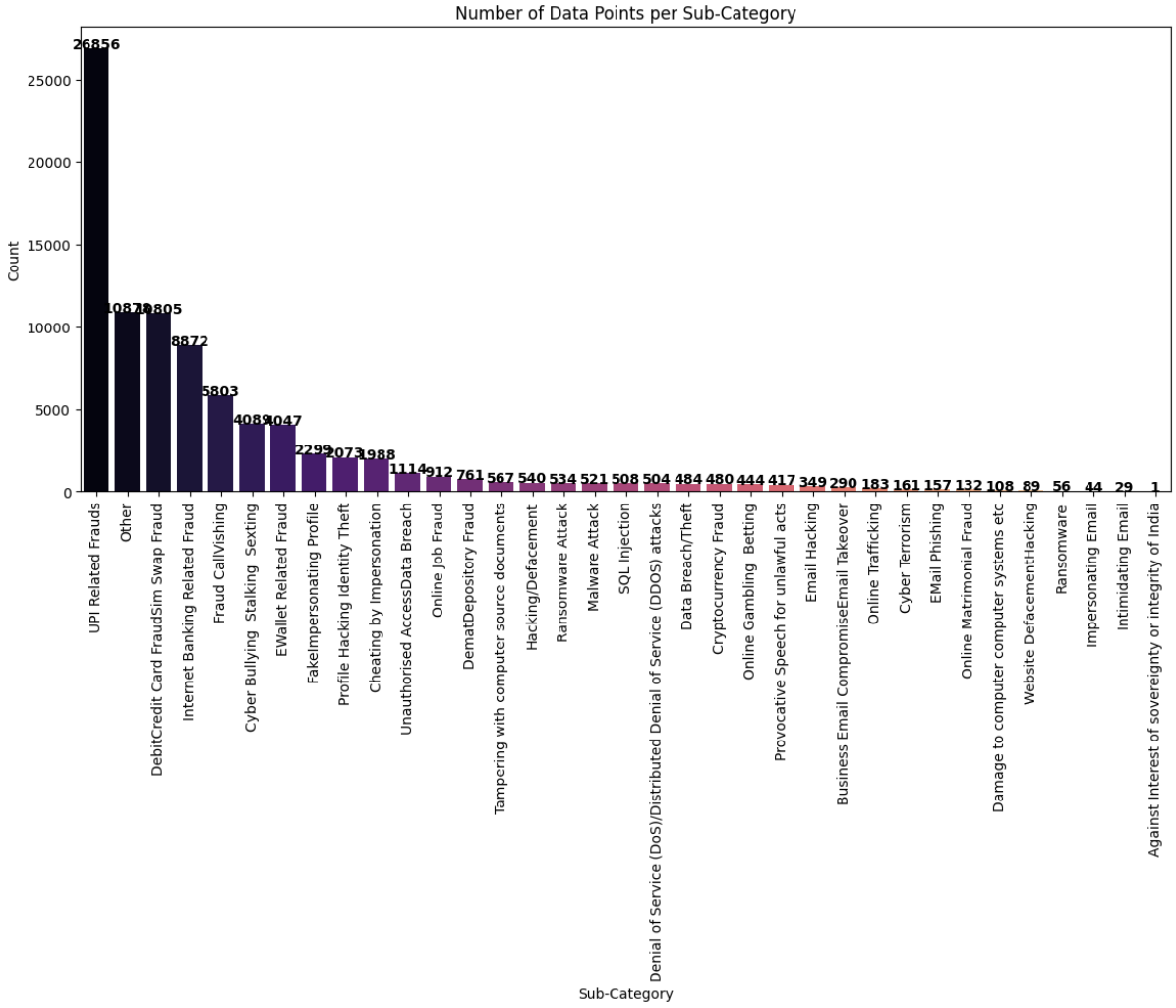
Number of categories having no sub categories: 4

Exploratory Data analysis

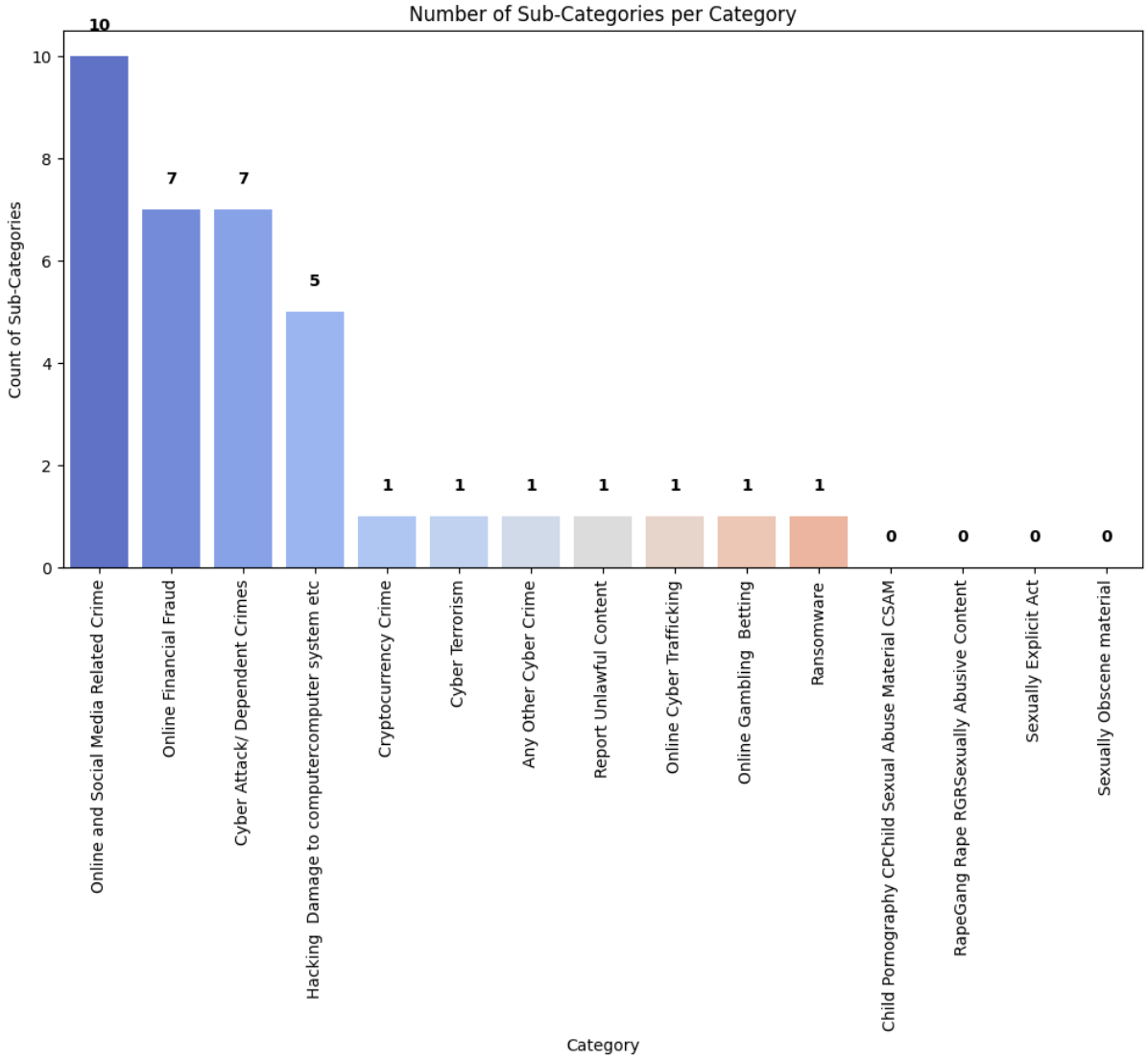
Number of data points for each category:



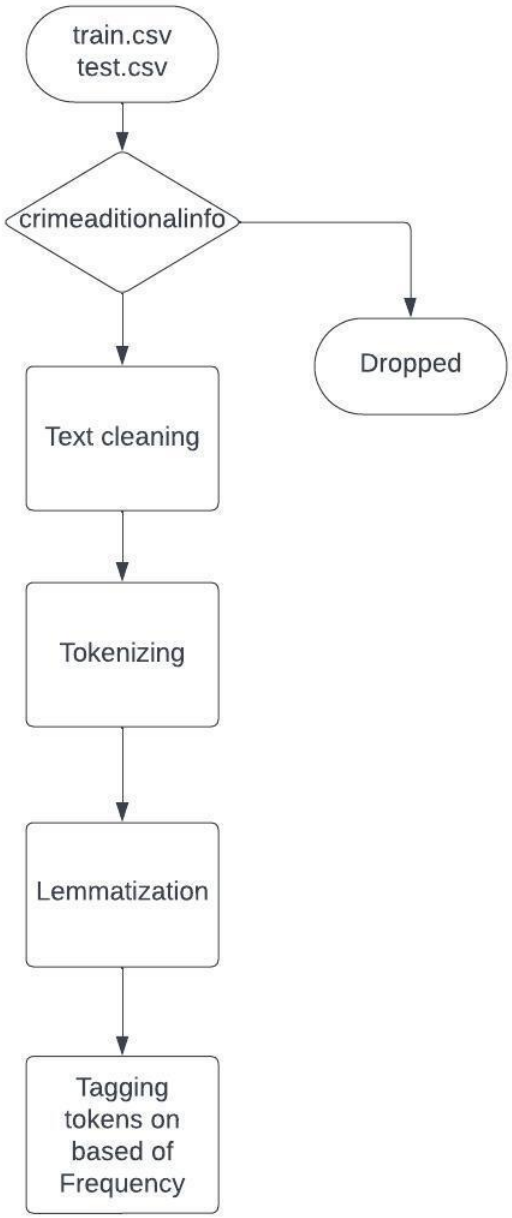
Number of data points for each sub category:



Number of sub category for each category:



Data Process of Cleaning



Exploratory Data Analysis

Generated unigrams, bigrams and Tri grams for each category

```python
# Recurring terms and jargon
for category in df_train['category'].unique():
    print(f"\nCategory: {category}")
    print(f"Top Unigrams: {[word[0][0] for word in top_unigrams[category]]}")
    print(f"Top Bigrams: {[f'{word[0][0]} {word[0][1]}' for word in top_bigrams[category]]}")
    print(f"Top Trigrams: {[f'{word[0][0]} {word[0][1]} {word[0][2]}' for word in top_trigrams[category]]}")
```

Python

```
Category: Online and Social Media Related Crime
Top Unigrams: ['number', 'video', 'call', 'account', 'money', 'please', 'facebook', 'identification', 'help', 'whatsapp']
Top Bigrams: ['video call', 'please help', 'social medium', 'asking money', 'facebook account', 'mobile number', 'whatsapp number', 'please take'
Top Trigrams: ['please please please', 'help help help', 'please take action', 'take necessary action', 'take strict action', 'video social mediu

Category: Online Financial Fraud
Top Unigrams: ['amount', 'account', 'bank', 'fraud', 'number', 'total', 'r', 'call', 'money', 'please']
Top Bigrams: ['total amount', 'take necessary', 'necessary action', 'amount please', 'account take', 'please hold', 'reverse total', 'hold revers
Top Trigrams: ['take necessary action', 'account take necessary', 'total amount please', 'reverse total amount', 'hold reverse total', 'please ho

Category: Online Gambling  Betting
Top Unigrams: ['money', 'please', 'number', 'sir', 'amount', 'account', 'app', 'r', 'help', 'call']
Top Bigrams: ['please help', 'money back', 'take action', 'many people', 'get money', 'cyber crime', 'sir please', 'mobile number', 'help sir', '
Top Trigrams: ['please help sir', 'get money back', 'please take action', 'please help get', 'help get money', 'money back please', 'card pan car

Category: RapeGang Rape RGRSexually Abusive Content
Top Unigrams: ['area', 'mall', 'activity', 'shamefull', 'place', 'involve', 'gariahat', 'sir', 'person', 'many']
Top Bigrams: ['shamefull activity', 'last year', 'involve shamefull', 'mall area', 'subhro saha', 'please help', 'sir person', 'respected sir', '
Top Trigrams: ['involve shamefull activity', 'respected sir serious', 'sir serious matter', 'serious matter want', 'matter want inform', 'want in
```

Using word to vec for vectorizing the tagged tokens

Filled sub category NA values with category names.

Different ML models accuracy and other parameters

Category

Algorithm Name	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	0.67	0.71	0.67	0.64
XGB	0.76	0.73	0.76	0.73
Light GBM	0.63	0.59	0.63	0.58
CatBoost	0.77	0.73	0.77	0.73
Gaussian Naïve Bayes	0.48	0.75	0.48	0.56
Random Forest	0.76	0.73	0.76	0.72

Sub Category

Algorithm Name	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	0.35	0.42	0.35	0.34
XGB	0.52	0.50	0.52	0.49
Light GBM	0.28	0.20	0.28	0.16
CatBoost	0.54	0.51	0.54	0.51
Gaussian Naïve Bayes	0.31	0.45	0.31	0.35
Random Forest	0.53	0.53	0.53	0.50

Bert Model Results

Category

Validation: 0.77

Testing: 0.76

Sub Category

Validation: 0.57

Testing: 0.57