CyberGuard AI Hackathon

M Sree Abhinav

Jagadesh

Analysis of Problem Statement & Dataset provided

1. The categories and subcategories in the datasets provided do not match the categories and sub-categories to be classified into for the Hackathon. For this reason, mere cleaning of the dataset provided and then training a model to make predictions will not be able to achieve acceptable accuracy.
2. There is a lot of class/label imbalance in the dataset provided.

Figure 1: Category distribution in Dataset1[test.csv]

Row Labels	**Count of crimeaditionalinfo**
Any Other Cyber Crime	3670
Child Pornography CPChild Sexual Abuse Material CSAM	123
Crime Against Women & Children	4
Cryptocurrency Crime	166
Cyber Attack/ Dependent Crimes	1261
Cyber Terrorism	52
Hacking Damage to computercomputer system etc	592
Online and Social Media Related Crime	4139
Online Cyber Trafficking	61
Online Financial Fraud	18890
Online Gambling Betting	134
Ransomware	18
RapeGang Rape RGRSexually Abusive Content	912
Sexually Explicit Act	535
Sexually Obscene material	665
Grand Total	**31222**

Figure 2: Category distribution in Dataset2[train.csv]

Row Labels	Count of crimeaditionalinfo
Any Other Cyber Crime	10877
Child Pornography CPChild Sexual Abuse Material CSAM	379
Cryptocurrency Crime	480
Cyber Attack/ Dependent Crimes	3608
Cyber Terrorism	161
Hacking Damage to computercomputer system etc	1710
Online and Social Media Related Crime	12138
Online Cyber Trafficking	183
Online Financial Fraud	57416
Online Gambling Betting	444
Ransomware	56
RapeGang Rape RGRSexually Abusive Content	2822
Report Unlawful Content	1
Sexually Explicit Act	1552
Sexually Obscene material	1838
Grand Total	**93665**

3. Other important demographic details of the victim/complainant would also help in classifying the complaint description, for example, crimes related to women and children.
4. It is possible that the complaint's category will overlap with multiple categories and subcategories.
5. 5. Many text complaints [crimeaditionalinfo] are transliterated from Hindi to English. This poses an additional challenge to any model or approach that considers the context of the words during the learning phase.
6. There are around 25 subcategories for which there is no corresponding data in either of the datasets.
7. Three subcategories in the one dataset do not have entries in the other dataset provided.

Date preprocessing

1. Fill missing sub-categories: In 6591 rows, the subcategories column had no entry. The value from the category column was copied into the subcategories column
2. Delete duplicates: There are 5998 Duplicate entries in the ‘crimeaditionalinfo’ column, which have been deleted.
3. Data cleaning 1: Removed all special characters, multiple spaces, line breaks, tab breaks, repetitive characters from ‘crimeaditionalinfo’ column and made all the characters lowercase.

4. Data cleaning 2: Removed all rows where the number of words in ‘crimeaditionalinfo’ was less than 5.
5. Removed rows that have crimeaditionalinfo as null/blank.
6. StopWords – Removed stopwords for traditional algorithms and not for Transformer algorithms[BERT] to maintain the context

Approach to text classification

This hackathon presents a classic text classification problem. Two approaches have been identified, though the team is being taken.

1. Traditional machine learning: Use of word vectorization like word2cev, countvectorize, and tfdif to generate vectors for the data in ‘crimeaditionalinfo’ and then running a traditional ML algorithm like Naïve Bayes, SVM, etc\
2. Transformer learning: Generation of vectors for the ‘crimeaditionalinfo’ column using tokenizers based on transformer models.

Three strategies have been developed, and accuracies, a confusion matrix, and an F1 score have been generated to establish a benchmark.

Count Vectorizer & Naïve Bayes

Vectors for the ' crimeaditionalinfo ' column were generated using the traditional technique of generating word embeddings using Count Vectorizer. Prior to this, since the embeddings were calculated based on the frequency of words, stopwords were removed, and lemmatization was performed.

```
# Vectorize text
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

A NaivesBayes algorithm was then trained based on the Train Dataset, and the model's accuracy scores were calculated on the test split.

```
Accuracy: 0.44726749760306805
                                                                    precision    recall  f1-score   support

                          Business Email CompromiseEmail Takeover       0.00      0.00      0.00        63
                                         Cheating by Impersonation       0.16      0.01      0.02       375
                  Child Pornography CPChild Sexual Abuse Material CSAM       0.00      0.00      0.00        88
                                              Cryptocurrency Fraud       0.80      0.04      0.07       105
                                      Cyber Bullying  Stalking  Sexting       0.38      0.67      0.49       793
                                                   Cyber Terrorism       0.00      0.00      0.00        23
                                Damage to computer computer systems etc       0.00      0.00      0.00        28
                                                 Data Breach/Theft       0.00      0.00      0.00        92
                                  DebitCredit Card FraudSim Swap Fraud       0.58      0.59      0.59      1903
                                          DematDepository Fraud       0.00      0.00      0.00       170
Denial of Service (DoS)/Distributed Denial of Service (DDOS) attacks       0.00      0.00      0.00       105
                                                     EMail Phishing       0.00      0.00      0.00        29
                                              EWallet Related Fraud       0.29      0.39      0.33       758
                                                      Email Hacking       0.00      0.00      0.00        55
                                          FakeImpersonating Profile       0.62      0.15      0.24       451
                                                  Fraud CallVishing       0.31      0.15      0.21      1107
                                                 Hacking/Defacement       0.16      0.55      0.25       108
                                                Impersonating Email       0.00      0.00      0.00        11
                                        Internet Banking Related Fraud       0.63      0.26      0.37      1541
                                                 Intimidating Email       0.00      0.00      0.00         8
                                                     Malware Attack       0.00      0.00      0.00       126
                                             Online Gambling  Betting       0.00      0.00      0.00        79
                                                   Online Job Fraud       0.33      0.01      0.01       171
                                            Online Matrimonial Fraud       0.00      0.00      0.00        22
                                                  Online Trafficking       0.00      0.00      0.00        27
                                                               Other       0.27      0.58      0.37      2106
                                         Profile Hacking Identity Theft       0.55      0.23      0.33       388
                                      Provocative Speech for unlawful acts       0.50      0.01      0.03        70
                                                          Ransomware       0.00      0.00      0.00        13
                                                   Ransomware Attack       0.14      0.29      0.19       102
                               RapeGang Rape RGRSexually Abusive Content       1.00      0.03      0.06        65
                                                       SQL Injection       0.25      0.01      0.02       112
                                               Sexually Explicit Act       0.00      0.00      0.00       302
                                            Sexually Obscene material       0.62      0.04      0.08       337
                                   Tampering with computer source documents       0.15      0.24      0.19       106
                                                  UPI Related Frauds       0.63      0.74      0.68      4588
                                         Unauthorised AccessData Breach       0.28      0.15      0.19       245
                                            Website DefacementHacking       0.00      0.00      0.00        16

                                                            accuracy                          0.45     16688
                                                           macro avg       0.23      0.14      0.12     16688
                                                        weighted avg       0.45      0.45      0.41     16688
```

An overall accuracy score of 0.45 was achieved. By analysing the classification report[confusion matrix], it was seen that the prediction of minority classes was very poor, and the weighted average accuracy was better than the macro accuracy.

TF-IDF & Naïve Bayes

Another traditional approach of using a TF-IDF vectorizer instead was tested.

```
# Vectorize text using TF-IDF
tfidf_vectorizer = TfidfVectorizer()
X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)
```

A NaiveBayes model was trained using the same training dataset, and accuracy was calculated on the Test split.

```
Accuracy: 0.331363986642692
                                                          precision    recall  f1-score   support

                     Business Email CompromiseEmail Takeover       0.00      0.00      0.00        53
                                 Cheating by Impersonation       0.00      0.00      0.00       385
                Child Pornography CPChild Sexual Abuse Material CSAM       0.00      0.00      0.00        69
                                     Cryptocurrency Fraud       0.00      0.00      0.00        84
                               Cyber Bullying  Stalking  Sexting       0.52      0.04      0.08       806
                                           Cyber Terrorism       0.00      0.00      0.00        31
                     Damage to computer computer systems etc       0.00      0.00      0.00        24
                                        Data Breach/Theft       0.00      0.00      0.00        89
                          DebitCredit Card FraudSim Swap Fraud       0.77      0.20      0.32      1713
                                  DematDepository Fraud       0.00      0.00      0.00       130
Denial of Service (DoS)/Distributed Denial of Service (DDOS) attacks       0.00      0.00      0.00       115
                                          EMail Phishing       0.00      0.00      0.00        31
                                    EWallet Related Fraud       0.75      0.11      0.20       754
                                           Email Hacking       0.00      0.00      0.00        67
                                FakeImpersonating Profile       0.00      0.00      0.00       460
                                       Fraud CallVishing       0.00      0.00      0.00      1093
                                     Hacking/Defacement       0.17      0.50      0.26       105
                                     Impersonating Email       0.00      0.00      0.00        11
                           Internet Banking Related Fraud       0.56      0.25      0.35      1390
                                     Intimidating Email       0.00      0.00      0.00         7
                                          Malware Attack       0.12      0.45      0.19        87
                                Online Gambling  Betting       0.00      0.00      0.00        96
                                        Online Job Fraud       0.00      0.00      0.00       157
                                Online Matrimonial Fraud       0.00      0.00      0.00        21
                                     Online Trafficking       0.00      0.00      0.00        31
                                                   Other       0.25      0.25      0.25      2196
                           Profile Hacking Identity Theft       0.00      0.00      0.00       400
                       Provocative Speech for unlawful acts       0.00      0.00      0.00        71
                                              Ransomware       0.00      0.00      0.00         7
                                       Ransomware Attack       0.00      0.00      0.00       119
                   RapeGang Rape RGRSexually Abusive Content       0.00      0.00      0.00        44
                                           SQL Injection       0.23      0.03      0.05       103
                                   Sexually Explicit Act       0.00      0.00      0.00       306
                                Sexually Obscene material       0.00      0.00      0.00       334
                    Tampering with computer source documents       0.12      0.07      0.09       110
                                      UPI Related Frauds       0.32      0.96      0.48      3841
                           Unauthorised AccessData Breach       0.00      0.00      0.00       209
                                Website DefacementHacking       0.00      0.00      0.00        23

                                                accuracy                          0.33     15572
                                               macro avg       0.10      0.08      0.06     15572
                                            weighted avg       0.32      0.33      0.24     15572
```

It was observed that both the accuracy and the macro average dropped. One reason for this could be the imbalance of data, which was directly affecting the vectors generated by the TFIDF vectorization. This is also confirmed by the fact that the F1 score of minority classes was very poor compared to the majority classes, as seen in the classification report.[confusion matrix].

As a possible remedy, random under and over-sampling of data was tried. As SMOTE was computationally expensive, traditional sampling methods were used. However, overall improvement was needed.

BERT

A transformer-based neural network model was developed to improve accuracy and develop a better model that can also understand the context to make predictions.

Since the task was text classification and not text generation, an encoder-only transformer model like BERT was zeroed in. The pre-trained model would be fine-tuned using the Train dataset provided. The BERT base model has 110 million parameters, and they can all be fine-tuned for the task at hand.

The preprocessing was done, but the stopwords were intentionally left in so as to maintain context. The tokenization was done using the BERT tokenizer, which is much more advanced than a traditional frequency-based tokenizer. The model was then trained on the Train dataset, and the accuracy was tested on the test dataset.

Accuracy: 0.5485
F1 Score: 0.5262

	precision	recall	f1-score	support
1	0.14	0.03	0.06	87
2	0.23	0.06	0.10	697
3	0.48	0.31	0.38	115
4	0.00	0.00	0.00	2
5	0.61	0.63	0.62	164
6	0.00	0.00	0.00	1
7	0.44	0.66	0.53	1304
8	0.00	0.00	0.00	51
9	0.00	0.00	0.00	34
10	0.12	0.03	0.05	171
11	0.71	0.75	0.73	3170
12	0.13	0.02	0.03	207
13	0.15	0.04	0.06	187
14	0.25	0.08	0.12	52
15	0.64	0.44	0.52	1267
16	0.39	0.37	0.38	128
17	0.47	0.44	0.46	734
18	0.32	0.31	0.32	1770
19	0.17	0.38	0.24	200
20	0.00	0.00	0.00	13
21	0.69	0.63	0.66	2643
22	0.00	0.00	0.00	11
23	0.22	0.15	0.18	170
32	0.15	0.13	0.14	186
33	0.00	0.00	0.00	97
34	0.13	0.24	0.17	167
35	0.00	0.00	0.00	1
36	0.25	0.02	0.03	516
37	0.30	0.24	0.26	646
38	0.18	0.13	0.15	194
39	0.68	0.82	0.74	7553
40	0.29	0.26	0.27	355
41	0.00	0.00	0.00	39
accuracy			0.55	27849
macro avg	0.26	0.22	0.22	27849
weighted avg	0.52	0.55	0.53	27849

An accuracy of .55 was achieved. The primary factor impacting the accuracy score is the imbalance in classes and the quality of the training data. Though the macro average improved drastically compared to traditional ML algorithms, it was still low.

Another strategy used a reduced sample size from the dataset for training. 100 samples from each label class were taken, and for those classes with less than 100 sample sizes, the entire data was taken. However, this did not result in a significant increase in accuracy scores.

Way forward

1. More data should be captured
2. The data should be accurately labelled. The categories should be realigned.
3. Prediction of the three most probable subcategories rather than a single category could help in better classification
4. Image classification on incident media files uploaded using CNN, RCNN, YOLO, multimodal LLMs, etc., in conjunction with text data to better classify complaints and take automated actions.

Conclusion

We evaluated the performance of a Large Language Model (LLM) like BERT and a traditional machine learning (ML) algorithm, Naive Bayes, on the given datasets. Our experiments corroborate the belief that LLMs often surpass traditional ML methods in sentiment analysis, spam SMS detection, and multi-label classification tasks. Moreover, the performance of LLMs can be further enhanced through fine-tuning strategies, making the fine-tuned models the top performers, as observed in our study.

Traditional ML and neural network (NN) approaches to text classification typically involve feature extraction, dimensionality reduction, and classifier selection, which can be complex, require domain expertise, and require considerable trial and error. In our study, we applied bag-of-words and TF-IDF methods before training the Naive Bayes classifier.

In contrast, LLMs simplify the text classification process by directly feeding data into the models and obtaining classification results. This straightforward approach eliminates the need for explicit feature extraction or dimensionality reduction, as LLMs inherently encode rich linguistic features through their deep contextual representations.