INDIAAI CYBERGUARD OFFICIAL ROUND 1 REPORT

Team Rugnite

1. Avanish Money Srivats
2. Tejeswar Pokuri
3. Komal Mathur
4. Shagnik Guha

Manipal Institute of Technology, Manipal

CONTENTS:

1. EXPLORATORY DATA ANALYSIS :

a) EDA Report and Analysis

Here are our findings from exploring the data:

1. The Class-wise distribution is heavily skewed hence any model building must take that into consideration.

2. In the test dataset:

   category: Non-null for all rows (object type).
   sub_category: Missing in 2,236 rows.
   crimeaditionalinfo: Missing in 7 rows.

Duplicates in the crimeaditionalinfo column: **2,443 (this includes those where the category repeats for multiple sub-categories)**

3. Some rows had empty info on the text side these were removed, it was also noticed that there were several duplicates of the same row, these were removed as well, although keeping them gives us better performance on the test dataset as it suffers from the same problem.
4. It was observed that out of all 14 categories, only 5 required further sub-category classification work as well, as the rest of the predictions are directly correlated to the category prediction(out of this Women and Children is also considered cos something).
5. Test labels not in the intersection of test and train: ['Crime Against Women & Children']
   Train labels not in the intersection of test and train: ['Report Unlawful Content']

   These categories were safely removed from the dataset, as a result, we can also observe that the following sub-categories:
   Test not in intersection: ['Cyber Blackmailing & Threatening', 'Computer Generated CSAM/CSEM', 'Sexual Harassment']
   Train not in the intersection: ['Against Interest of sovereignty or integrity of India']

b) Dataset Overview

TRAIN:

category	
Online Financial Fraud	57434
Online and Social Media Related Crime	12140
Any Other Cyber Crime	10878
Cyber Attack/ Dependent Crimes	3608
RapeGang Rape RGRSexually Abusive Content	2822
Sexually Obscene material	1838
Hacking Damage to computercomputer system etc	1710
Sexually Explicit Act	1552
Cryptocurrency Crime	480
Online Gambling Betting	444
Child Pornography CPChild Sexual Abuse Material CSAM	379
Online Cyber Trafficking	183
Cyber Terrorism	161
Ransomware	56
Report Unlawful Content	1

Table 1: training data label split

TEST:

category	
Online Financial Fraud	18896
Online and Social Media Related Crime	4139
Any Other Cyber Crime	3670
Cyber Attack/ Dependent Crimes	1261
RapeGang Rape RGRSexually Abusive Content	912
Sexually Obscene material	666
Hacking Damage to computercomputer system etc	592
Sexually Explicit Act	535
Cryptocurrency Crime	166
Online Gambling Betting	134
Child Pornography CPChild Sexual Abuse Material CSAM	123
Online Cyber Trafficking	61
Cyber Terrorism	52
Ransomware	18
Crime Against Women & Children	4

Table 2: test data label split

c) Visualisations

	category	sub_category
0	Any Other Cyber Crime	1
1	Child Pornography CPChild Sexual Abuse Materia...	0
2	Cryptocurrency Crime	1
3	Cyber Attack/ Dependent Crimes	7
4	Cyber Terrorism	1
5	Hacking Damage to computercomputer system etc	5
6	Online Cyber Trafficking	1
7	Online Financial Fraud	7
8	Online Gambling Betting	1
9	Online and Social Media Related Crime	10
10	Ransomware	1
11	RapeGang Rape RGRSexually Abusive Content	0
12	Report Unlawful Content	1
13	Sexually Explicit Act	0
14	Sexually Obscene material	0

Table 3: No of sub-categories per category

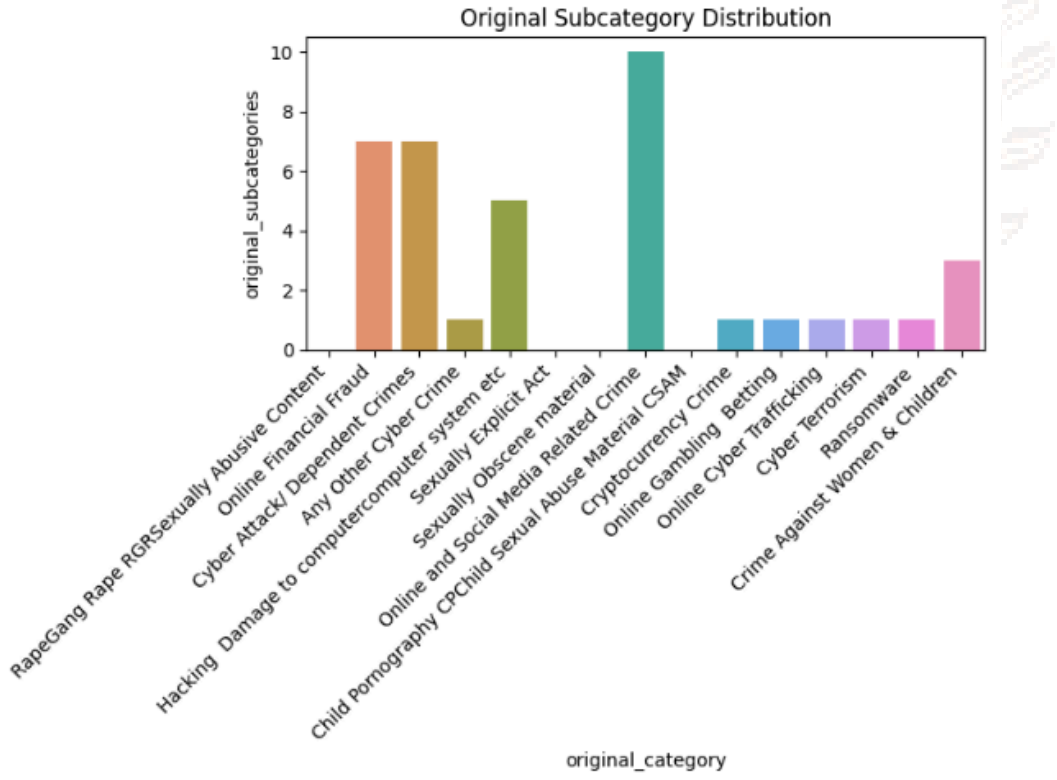


Figure 1: Figure illustrating the distribution among categories

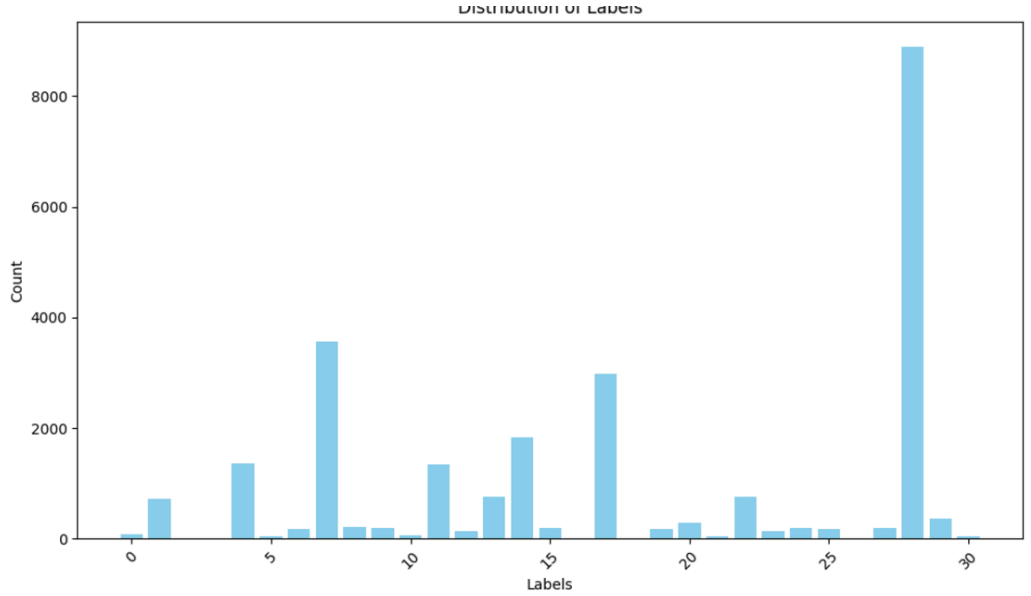


Figure 2: Figure illustrating the class imbalance in the sub-categories

Subcategory label distribution

	label	count	sub_category
0	28	8890	UPI Related Frauds
1	7	3556	DebitCredit Card FraudSim Swap Fraud
2	17	2973	Internet Banking Related Fraud
3	14	1827	Fraud CallVishing
4	4	1366	Cyber Bullying Stalking Sexting
5	11	1338	EWallet Related Fraud
6	13	763	FakeImpersonating Profile
7	22	751	Profile Hacking Identity Theft
8	1	719	Cheating by Impersonation
9	29	370	Unauthorised AccessData Breach
10	20	294	Online Job Fraud
11	8	222	DematDepository Fraud
12	15	200	Hacking/Defacement
13	27	194	Tampering with computer source documents
14	9	187	Denial of Service (DoS)/Distributed Denial of ...
15	24	186	Ransomware Attack
16	6	171	Data Breach/Theft
17	19	170	Malware Attack
18	25	167	SQL Injection
19	12	130	Email Hacking
20	23	130	Provocative Speech for unlawful acts
21	0	90	Business Email CompromiseEmail Takeover
22	10	54	EMail Phishing
23	5	39	Damage to computer computer systems etc
24	30	39	Website DefacementHacking
25	21	38	Online Matrimonial Fraud
26	16	13	Impersonating Email
27	18	11	Intimidating Email
28	2	2	Computer Generated CSAM/CSEM
29	3	1	Cyber Blackmailing & Threatening
30	26	1	Sexual Harassment

Table 4: Subcategories and their specific counts

2. Data Pre-Processing Techniques:

a. Data Augmentation:

Data Augmentation Tactics with Llama:

To undercut some of the data difficulties, mainly with class imbalance, we decided to leverage LLMs to create some more datasets for the sparsely populated classes. This was done with a well-crafted prompt that prompted it to create more data, given a particular row, by rephrasing sentences, thus retaining original content but also increasing dataset size. The results are shown below in the handling imbalance section.

b. Data Cleaning

All the text was preprocessed in the following manner:

1. Removal of empty rows, ie. where the text row is empty
2. Lemmatisation, ie. breaking down words into their root form
3. Stopword removal, ie. removal of commonly occurring texts

c. Feature Engineering

Find no. of new lines, and tab characters, tested on other regex characters as well, and find correlations between the count of the regex characters.

d. Handling Imbalance

category	
Online Financial Fraud	57434
Online and Social Media Related Crime	12140
Any Other Cyber Crime	10878
Cyber Attack/ Dependent Crimes	3608
Sexually Explicit Act	3000
RapeGang Rape RGRSexually Abusive Content	2822
Sexually Obscene material	1838
Hacking Damage to computercomputer system etc	1710
Online Gambling Betting	1000
Cyber Terrorism	500
Online Cyber Trafficking	500
Cryptocurrency Crime	480
Child Pornography CPChild Sexual Abuse Material CSAM	379
Ransomware	300
Report Unlawful Content	1

Table 5: After data augmentation (explained below)

3. Models and Results

In this section, we present our approach and results for category and sub-category classification. Initially, we experimented with baseline machine learning models commonly used for classification tasks. As anticipated, these models yielded low accuracy. To improve performance, we employed NLP-based text classification models, including **HingRoBERTa**, **BERT Base-Uncased**, and **DistilBERT Hinglish**.

While these models demonstrated decent accuracy, they struggled to handle the class imbalance effectively. Among them, **HingRoBERTa** outperformed the others in terms of accuracy. Consequently, we developed a custom pipeline leveraging **HingRoBERTa** to achieve improved classification accuracy.

A)NLP-based text Classification Models

We have decided to use **HingRoBERTa, BERT Base-Uncased and DistliBERT Hinglish**

Hinglish RoBERTa - Hinglish RoBERTa is a fine-tuned version of RoBERTa tailored for the Hinglish (Hindi-English) code-mixed text. It leverages robust pretraining and large-scale token masking to understand bilingual context effectively. In our experiments, it demonstrated superior performance compared to other models due to its adaptability to the Hinglish dataset.

BERT Base-Uncased - BERT base-uncased is a general-purpose pre-trained transformer model that uses a bidirectional encoder for language understanding. It operates on lowercase English text, which may limit its effectiveness with code-mixed or multilingual data. While it provided moderate accuracy, its inability to fully understand the Hinglish context reduced its performance.

DistliBERT Hinglish - DistilBERT Hinglish is a lightweight version of BERT, optimized for efficiency and speed while retaining good performance for Hinglish text. It uses fewer parameters, making it faster and more memory-efficient compared to other models. In our experiments, it provided comparable results but fell short in accuracy relative to Hinglish RoBERTa. Its reduced capacity likely contributed to limitations in learning from the class-imbalanced data.

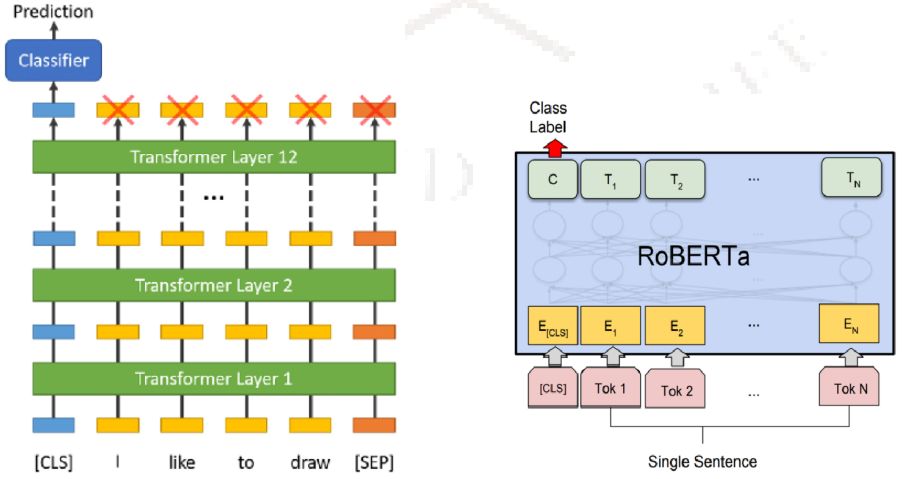


Figure 3: BERT and RoBERTa architecture summarized

1) Category

We have directly trained the above-mentioned 3 models for the 15 categories, you can see the results below

Hinglish RoBERTa

	precision	recall	f1-score	support
Any Other Cyber Crime	0.50	0.29	0.37	3670
Child Pornography CPChild Sexual Abuse Material CSAM	0.57	0.36	0.44	123
Cryptocurrency Crime	0.59	0.61	0.60	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	0.39	0.35	0.37	592
Online Cyber Trafficking	0.00	0.00	0.00	61
Online Financial Fraud	0.83	0.95	0.88	18890
Online Gambling Betting	0.50	0.03	0.06	134
Online and Social Media Related Crime	0.59	0.60	0.60	4139
Ransomware	0.00	0.00	0.00	18
RapeGang Rape RGRSexually Abusive Content	0.98	0.91	0.95	912
Report Unlawful Content	0.00	0.00	0.00	535
Sexually Explicit Act	0.39	0.17	0.24	665
accuracy			0.77	31218
macro avg	0.45	0.38	0.39	31218
weighted avg	0.73	0.77	0.74	31218

Table 6: RoBERTa classification report

DistilBERT Hinglish

	precision	recall	f1-score	support
Any Other Cyber Crime	0.43	0.16	0.23	3670
Child Pornography CPChild Sexual Abuse Material CSAM	0.52	0.26	0.35	123
Cryptocurrency Crime	0.38	0.10	0.15	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	0.36	0.19	0.25	592
Online Cyber Trafficking	0.00	0.00	0.00	61
Online Financial Fraud	0.79	0.96	0.87	18890
Online Gambling Betting	0.00	0.00	0.00	134
Online and Social Media Related Crime	0.56	0.58	0.57	4139
Ransomware	0.00	0.00	0.00	18
RapeGang Rape RGRSexually Abusive Content	1.00	0.90	0.95	912
Report Unlawful Content	0.00	0.00	0.00	535
Sexually Explicit Act	0.36	0.07	0.12	665
accuracy			0.75	31218
macro avg	0.39	0.30	0.32	31218
weighted avg	0.69	0.75	0.71	31218

Table 7: DistilBERT classification report

BERT Base-Uncased

	precision	recall	f1-score	support
Any Other Cyber Crime	0.49	0.28	0.35	3670
Child Pornography CPChild Sexual Abuse Material CSAM	0.55	0.28	0.37	123
Cryptocurrency Crime	0.58	0.50	0.54	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	0.37	0.35	0.36	592
Online Cyber Trafficking	0.00	0.00	0.00	61
Online Financial Fraud	0.83	0.95	0.88	18890
Online Gambling Betting	0.50	0.01	0.01	134
Online and Social Media Related Crime	0.59	0.60	0.59	4139
Ransomware	0.00	0.00	0.00	18
RapeGang Rape RGRSexually Abusive Content	0.98	0.91	0.95	912
Report Unlawful Content	0.00	0.00	0.00	535
Sexually Explicit Act	0.36	0.17	0.23	665
accuracy			0.77	31218
macro avg	0.45	0.36	0.38	31218
weighted avg	0.73	0.77	0.74	31218

Table 8: bert-base-uncased classification report

Clearly the Hinglish RoBERTa is better than the other models, so we are using this model to increase accuracy and using a custom pipeline to overcome the huge data imbalance.

2) Sub category

hing-roberta

	precision	recall	f1-score	support
Business Email CompromiseEmail Takeover	0.00	0.00	0.00	90
Cheating by Impersonation	0.56	0.66	0.61	719
Computer Generated CSAM/CSEM	0.00	0.00	0.00	2
Cyber Blackmailing & Threatening	0.00	0.00	0.00	1
Cyber Bullying Stalking Sexting	0.72	0.89	0.79	1366
Damage to computer computer systems etc	0.00	0.00	0.00	39
Data Breach/Theft	0.00	0.00	0.00	171
DebitCredit Card FraudSim Swap Fraud	0.83	0.81	0.82	3556
DematDepository Fraud	0.58	0.11	0.19	222
Denial of Service (DoS)/Distributed Denial of Service (DDOS) attacks	0.15	1.00	0.26	187
EMail Phishing	0.00	0.00	0.00	54
EWallet Related Fraud	0.74	0.52	0.61	1338
Email Hacking	0.60	0.76	0.67	130
FakeImpersonating Profile	0.68	0.49	0.57	763
Fraud CallVishing	0.72	0.64	0.68	1827
Hacking/Defacement	0.00	0.00	0.00	200
Impersonating Email	0.00	0.00	0.00	13
Internet Banking Related Fraud	0.81	0.68	0.74	2973
Intimidating Email	0.00	0.00	0.00	11
Malware Attack	0.00	0.00	0.00	170
Online Job Fraud	0.88	0.83	0.86	294
Online Matrimonial Fraud	0.00	0.00	0.00	38
Profile Hacking Identity Theft	0.70	0.70	0.70	751
Provocative Speech for unlawful acts	0.00	0.00	0.00	130
Ransomware Attack	0.00	0.00	0.00	186
SQL Injection	0.00	0.00	0.00	167
Sexual Harassment	0.00	0.00	0.00	1
Tampering with computer source documents	0.00	0.00	0.00	194
UPI Related Frauds	0.80	0.93	0.86	8890
Unauthorised AccessData Breach	0.77	0.89	0.82	370
Website DefacementHacking	0.00	0.00	0.00	39
accuracy			0.74	24892
macro avg	0.31	0.32	0.30	24892
weighted avg	0.72	0.74	0.73	24892

Table 9: hing-roberta classification report on subcategories

Our current approach involved using the best-performing model on the previous dataset and training the base model again for a different purpose, this work can be further enhanced by using the weights learnt from the category classification task, even though the classes are different as the corpora is the same, we expect good results, and is part of our future work. This process is known as transfer learning.
The major setback faced in this task, remains the same as well, being that the low support values cause the model to never learn some class features, despite that the overall accuracy still falls to 0.74.
We also made efforts to use the previous pipeline used in categories, involving 2 models, but it might be a more reliable approach to first classify for the category and create an individual model, for each label. Moreover, the task is far simpler than the category task, as most categories either have 1 or 0 sub-categories as detailed in the EDA section, hence proper models if required would only be needed for 5 proper sub-classification tasks.

B)Our own pipeline

We observed that some classes were consistently overlooked by the model, a phenomenon known as **mode collapse**, which is a common issue in deep learning models, especially when handling a large number of classes and imbalanced datasets. To address this, we designed a robust pipeline.

The pipeline operates as follows:

1. As identified during the Exploratory Data Analysis (EDA), the class for "online financial fraud" dominates the dataset, accounting for more than half of the instances.
2. To mitigate the imbalance, the first model in the pipeline is trained to classify whether an instance belongs to the "online financial fraud" class or not.
3. For instances classified as not belonging to this dominant class, a second model is deployed to predict the specific class among the remaining categories.

For this setup, we utilized the **Hinglish RoBERTa** model, as it delivered the best performance in category classification. By employing this two-step pipeline, we achieved a significant improvement, with macro accuracy increasing by 66%. This approach has proven to be highly effective for addressing the challenges posed by a multiclass imbalanced dataset.

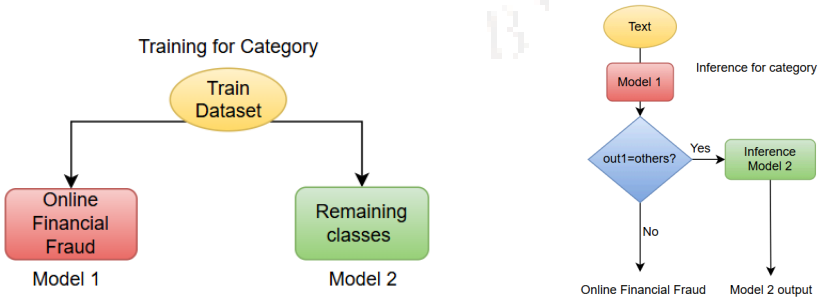


Figure 4: Our custom pipeline

	precision	recall	f1-score	support
Online Financial Fraud	0.86	0.87	0.87	18890
Any Other Cyber Crime	0.65	0.68	0.66	3670
Child Pornography CPChild Sexual Abuse Material	0.52	0.36	0.42	123
Cryptocurrency Crime	0.68	0.73	0.71	166
Cyber Attack/ Dependent Crimes	1.00	1.00	1.00	1261
Cyber Terrorism	0.50	0.55	0.52	52
Hacking Damage to computercomputer system etc	0.40	0.40	0.40	592
Online Cyber Trafficking	0.31	0.34	0.32	61
Online Gambling Betting	0.42	0.46	0.44	134
Online and Social Media Related Crime	0.61	0.65	0.63	4139
Ransomware	0.36	0.22	0.28	18
RapeGang Rape RGRSexually Abusive Content	0.94	0.92	0.93	912
Sexually Explicit Act	0.25	0.16	0.19	535
Sexually Obscene material	0.38	0.30	0.33	665
Crime Against Women & Children	0.00	0.00	0.00	4
accuracy			0.78	31222
macro avg	0.52	0.51	0.52	31222
weighted avg	0.75	0.76	0.75	31222

Table 10: Custom pipeline classification report

4. Conclusion:

After intense testing of various approaches ranging from basic ML models combined with TF-IDF to complex large embedding models and tokenizers such as BERT we have the following conclusions to present in terms of the most effective models to solve the given dataset.

The best performing model for the task of category classification was the - hing-roberta 78% using our approach detailed above, with no mode collapse.

While there is more scope of work for the sub-category classification task, our current model achieves results that nears the classification task (74%, with minor mode collapse), meaning it is minimally overfitting, ie. it is the most concrete proof we have to present that the model is generalisable.

5. Future Work:

a) Data Processing Techniques

1. Implement advanced text augmentation techniques using multiple LLMs to better handle class imbalance, as we have already explored this avenue with LLAMA we will explore better prompting techniques and other models as well.
2. Explore cross-lingual data augmentation using parallel corpora and intermixed corpora and explore results.
3. Implement advanced cleaning techniques for code-mixed text, like SMOTE, back translation, paraphrasing etc.

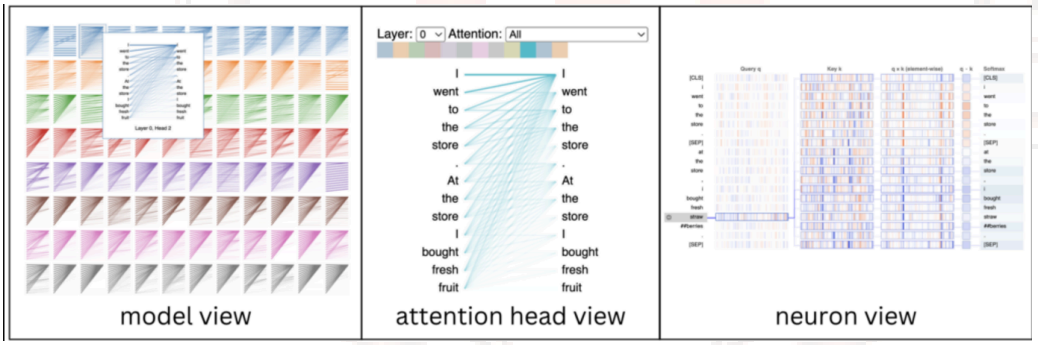
b) Model Avenues to Explore

1. Implement hierarchical classification approaches to better handle category-subcategory relationships
2. Explore ensemble methods combining multiple specialized models for different crime types
3. Investigate few-shot learning approaches for rare crime categories
4. Develop multilingual models specifically trained for Indian language variations
5. Research transfer learning from larger cybersecurity datasets to improve feature extraction

The focus is on improving model performance for underrepresented classes while maintaining accuracy on majority classes, basically preventing mode collapse, without overfitting. Additionally, investigating domain adaptation techniques could help better handle emerging cybercrime.

C) Explainable AI

Another avenue we consider exploring is XAI or Explainable AI, which makes our models understandable, interprets where the model fails, and finds the exact reason why the model performs well. Our current ideas involve using LIME, SHAP, and BERTviz to achieve this purpose.



—---------- THANK YOU----------