INDIAAI CYBERGUARD HACKATHON

PROJECT REPORT

TEAM DETAILS:

Team Name: Hacktastic Coders

Team Members:

1. Aadithya S – Sri Sairam Institute of Technology [Team Leader]

2. Dr. Su. Suganthi – Sri Sairam Institute of Technology

3. Mrs. Illakiya - Sri Sairam Institute of Technology

4. Nivetha A - Sri Sairam Institute of Technology

5. Sanjitha S - Sri Sairam Institute of Technology

6. Samsthana K L - Sri Sairam Institute of Technology

TABLE OF CONTENTS

Chapter No.	**Topic**
	ABSTRACT
1	INTRODUCTION
	1.1 Problem Statement
2	DATA COLLECTION AND PROCESSING
	2.1 Dataset
	2.2 Data Preprocessing
	2.3 Exploratory Data Analysis (EDA)
3	DETAILED SOLUTION
	3.1 Model Selection
	3.2 Model Training and Fine-Tuning
	3.3 Model Evaluation and Analysis
	3.4 Significant Analysis from NLP analysis
4	CONCLUSION
	4.1 Conclusion
	4.2 Future Scope
	REFERENCES
	PLAGIARISM DECLARATION

ABSTRACT

Effectively managing and categorizing complaints is a critical challenge for organizations seeking to address fraud detection and streamline grievance handling. This project leverages DistilBERT, a cutting-edge transformer-based NLP model, to create a high-performance text classification system capable of categorizing complaints based on parameters such as the victim, type of fraud, and related factors. The process begins with meticulous text preprocessing, including tokenization, stop-word removal, and text cleaning, to ensure the data is optimized for analysis. DistilBERT is fine-tuned on the processed dataset to deliver precise and efficient classification. The model's reliability and effectiveness are validated through performance metrics such as accuracy, precision, recall, and F1-score. The resulting solution is lightweight, scalable, and capable of processing large volumes of data, enabling organizations to uncover fraudulent patterns, prioritize critical cases, and enhance operational efficiency.

Keywords: DistilBERT, complaint categorization, fraud detection, text preprocessing, transformer model, text classification, accuracy, precision, recall, F1-score.

CHAPTER 1

INTRODUCTION

1.1 PROBLEM STATEMENT:

Cyber crimes and scams have escalated significantly with the rise of digital platforms, causing billions of dollars in losses globally. These scams leverage deceptive tactics such as phishing emails, fraudulent investments, and fake customer service impersonations to exploit victims. Detecting and mitigating these scams in real time has become critical for user protection and financial security.

This project leverages Natural Language Processing (NLP) with DistilBERT to classify scam-related content, providing a scalable, efficient, and accurate solution for detecting scams in textual data**.**

CHAPTER 2

DATA COLLECTION AND PROCESSING

2.1 DATASET:

- The train and test dataset which was available in the official website of the hackathon was utilized.
- The datasets together contained around 1.56 lakhs rows with category, subcategory and crime information labels.

2.2 DATA PREPROCESSING:

- At first all the redundant and missing values were removed from the dataset
- Then tokenization and stop-words removal were done.
- At last a new label named " Cleaned_crime_info" was created which consisted of summarized and apt content of the crime information that was provided.

2.3 EXPLORATORY DATA ANALYSIS (EDA):

- EDA was performed on the train dataset to explore the data given.
- Several visualization techniques were employed to understand data effectively.

2.3.1 PROCEDURE:

- At first the dataset was loaded.
- The overall category and subcategory distribution was visualized using a pie-chart.

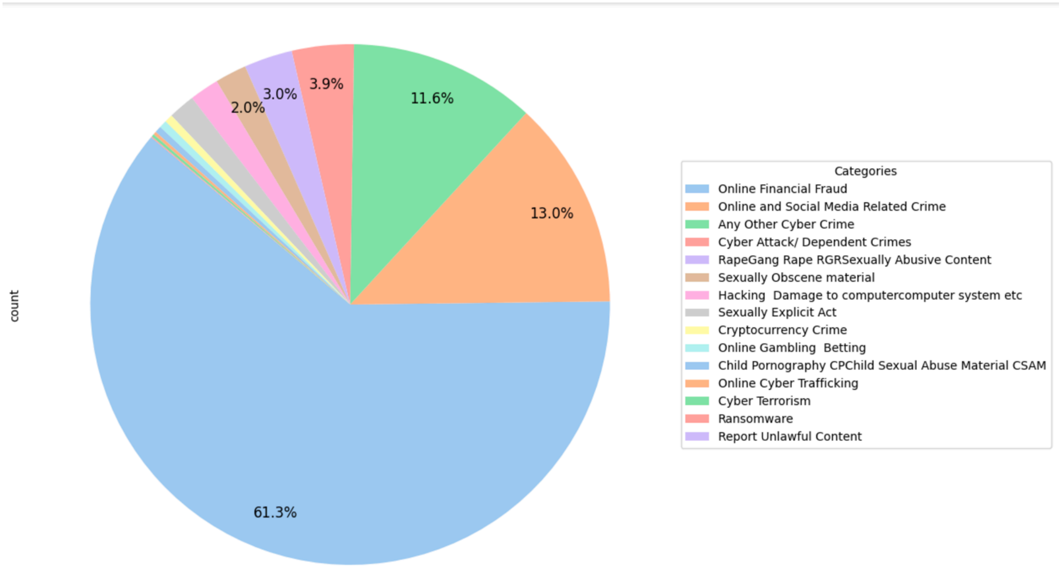


Fig 1. Visualization of Category Distribution

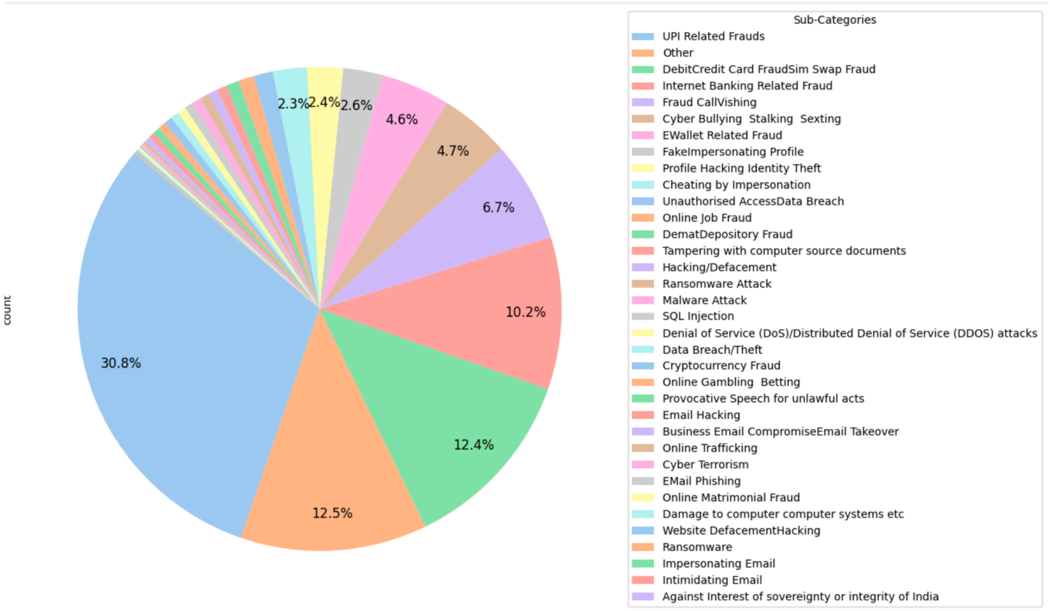


Fig 2. Visualization of Sub-Category Distribution

- Then the count and percentage of sub-categories within each main category were visualized using bar charts.

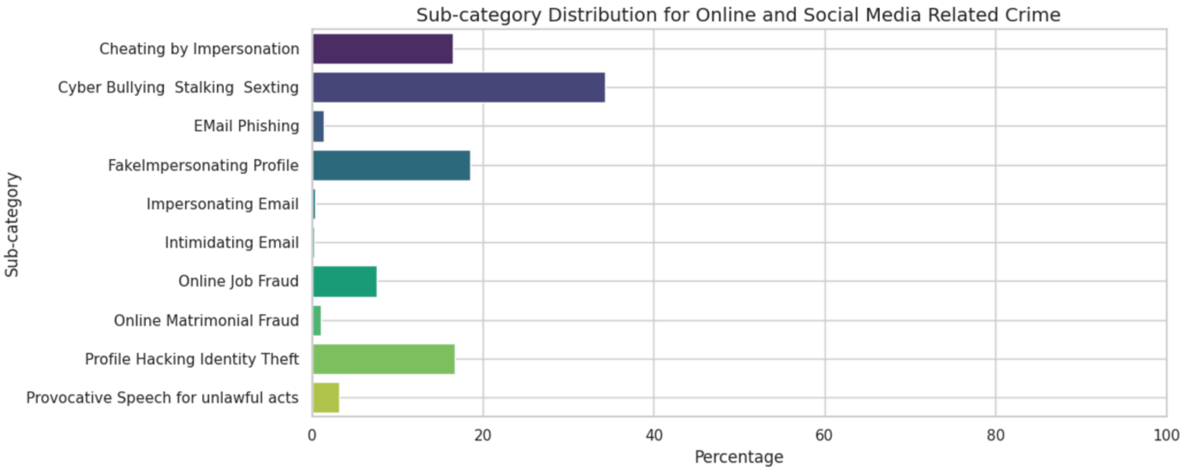


Fig 3. Sub-category Distribution for Online and Social Media Related Crime

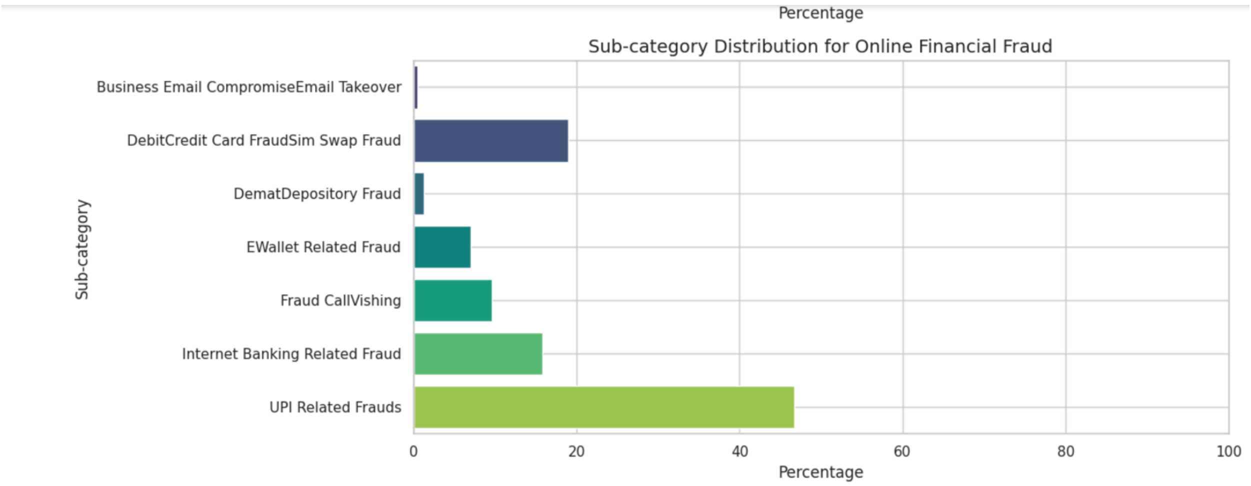


Fig 4. Sub-category Distribution for Online Financial Fraud

2.3.2 FINDINGS:

- There are total of 15 main categories and 35 sub-categories in the train dataset
- The highest number of complaints are under the “Online Financial Fraud” category with 61.3%.
- The highest number of complaints are under the “UPI Related frauds” with 30.8%.
- The minimum number of complaints registered under “unlawful content” with less than 1%.

CHAPTER 3

DETAILED SOLUTION

3.1 MODEL SELECTION:

- The model selection process evaluated various machine learning and transformer-based models for complaint categorization.
- Initially, ensemble methods like XGBoost and AdaBoost were tested but fell short in capturing the complex contextual nuances of textual data.
- To address this, transformer-based models such as BERT, mDistilBERT, and DistilBERT were explored.
- BERT showed strong contextual understanding but required high computational resources, limiting its scalability. mDistilBERT improved efficiency but did not significantly enhance accuracy.
- DistilBERT emerged as the best performer, achieving the highest accuracy of 0.7718, combining lightweight architecture with robust contextual learning.
- Its ability to balance computational efficiency and classification accuracy made it the most suitable for real-world applications.
- This evaluation concluded with DistilBERT as the optimal model, offering an accurate, efficient, and scalable solution for categorizing complaints effectively and addressing operational needs.

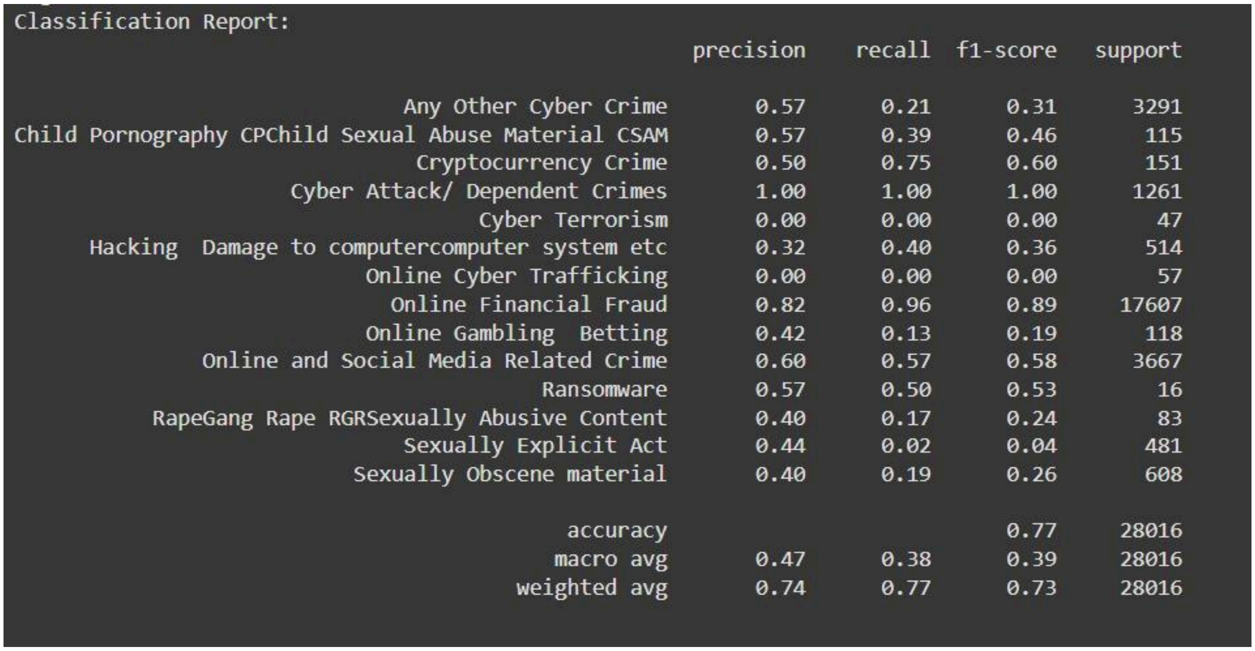
3.2 MODEL TRAINING AND FINE-TUNING:

- We have trained the Distilbert model using transformers' training arguments with the following parameters:
    - Learning rate: 5e-5
    - Epochs : 6
    - Weight- decay : 0.01
- DistilBERT was fine-tuned on labeled datasets using classification heads to distinguish between scam and non-scam messages.
- Hyperparameters optimized for domain-specific tasks.

3.3 MODEL EVALUATION AND ANALYSIS:

- The model was tested using various metrics like accuracy, precision score, recall, f1 scores,etc.
- The model works perfectly for labels with large number entries but it has decreased efficiency for labels with less number of entries.

Evaluation metric	Scores(in percentage)
Accuracy	77
Precision	74
Recall	77
F1 score	73



```
Classification Report:
                                                      precision    recall  f1-score   support

                               Any Other Cyber Crime       0.57      0.21      0.31      3291
Child Pornography CPChild Sexual Abuse Material CSAM       0.57      0.39      0.46       115
                                Cryptocurrency Crime       0.50      0.75      0.60       151
                      Cyber Attack/ Dependent Crimes       1.00      1.00      1.00      1261
                                     Cyber Terrorism       0.00      0.00      0.00        47
      Hacking  Damage to computercomputer system etc       0.32      0.40      0.36       514
                            Online Cyber Trafficking       0.00      0.00      0.00        57
                              Online Financial Fraud       0.82      0.96      0.89     17607
                            Online Gambling  Betting       0.42      0.13      0.19       118
               Online and Social Media Related Crime       0.60      0.57      0.58      3667
                                          Ransomware       0.57      0.50      0.53        16
           RapeGang Rape RGRSexually Abusive Content       0.40      0.17      0.24        83
                               Sexually Explicit Act       0.44      0.02      0.04       481
                           Sexually Obscene material       0.40      0.19      0.26       608

                                            accuracy                           0.77     28016
                                           macro avg       0.47      0.38      0.39     28016
                                        weighted avg       0.74      0.77      0.73     28016
```

Fig 5. Classification report of the model

3.4 SIGNIFICANT FINDINGS FROM NLP ANALYSIS:

- Fine-tuning DistilBERT on the dataset enabled the model to achieve high accuracy in categorizing complaints based on parameters such as victim type, nature of fraud, and other relevant factors.
- Despite the potential imbalance in complaint categories, the model demonstrated robust performance, with balanced precision and recall across categories, ensuring no significant bias in classification.
- Text preprocessing, including tokenization, stop-word removal, and text cleaning, played a critical role in enhancing model performance by reducing noise and standardizing input data.
- DistilBERT's ability to capture the context and nuances of text significantly improved classification accuracy, especially for complaints with complex or ambiguous phrasing.
- The lightweight nature of DistilBERT ensured faster training and inference times compared to larger models while maintaining high classification accuracy, making it suitable for real-world applications with large-scale complaint data.
- The analysis revealed recurring patterns and trends in fraud-related complaints, such as common keywords, phrases, or behaviors associated with specific fraud types, enabling actionable insights for preventive measures.
- Performance metrics—accuracy, precision, recall, and F1-score—indicated that the model consistently delivered reliable results across diverse complaint categories, establishing its effectiveness for deployment.

CHAPTER 4

CONCLUSION

4.1 CONCLUSION:

This project successfully developed an efficient and accurate complaint categorization system using DistilBERT, a state-of-the-art transformer-based NLP model. Through comprehensive text preprocessing and rigorous model evaluation, DistilBERT emerged as the optimal choice, achieving an accuracy of **0.7718**. Its ability to understand complex contextual nuances and maintain computational efficiency made it ideal for real-world deployment. The model effectively categorized complaints based on parameters such as victim type and fraud category, offering a scalable solution for organizations to enhance fraud detection and grievance handling processes. This work highlights the transformative role of advanced NLP techniques in automating and streamlining text classification tasks.

4.2 FUTURE SCOPE:

- Adding more contextual metadata, like locations, timestamps, or categories, may increase the accuracy of classification.
- Making use of the classified output for fraud trend analysis and predictive analytics may yield more in-depth information for making decisions.

REFERENCES:

1. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019)
   BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
   Presented at NAACL-HLT. Available at https://arxiv.org/abs/1810.04805.

2. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019)
   DistilBERT, a distilled version of BERT: Smaller, faster, cheaper, and lighter.
   Available at https://arxiv.org/abs/1910.01108.

3. Hugging Face Documentation

   - Official documentation for the Transformers and Datasets libraries.
     Website: https://huggingface.co/docs.

4. Scikit-learn Documentation

   - Evaluation metrics such as precision, recall, and F1-score.
     Website: https://scikit-learn.org/.

PLAGIARISM DECLARATION:

This project report is based on original analysis and research conducted using publicly available datasets and open-source libraries. All referenced works, tools, and datasets have been duly acknowledged to ensure transparency and integrity. No uncredited materials have been used in this project

The implementation plan, evaluation, and findings represent unique contributions tailored to the problem of online financial scams and are presented in accordance with ethical research and academic standards.