MINISTRY OF ELECTRONICS & INFORMATION TECHNOLOGY GOVERNMENT OF INDIA
Digital India
INDIAai
Indian Cyber Crime Coordination Centre
IndiaAI CyberGuard
AI Hackathon

**Project Report**

**Submission for IndiaAI CyberGuard AI Hackathon**

**Team name: Team Raksha**

**Members: Satvik Mishra, Kshitij Thareja**

**Date of Submission: November 22, 2024**

**Table of contents**

# Part 1: Summary of Work

## 1.1. Problem Statement: Development of an NLP Model for Text Analytics and Classification.

1.2. Solution Overview: For the current stage of the hackathon, the problem statement is centered around building an NLP model to classify online cybercrime related complaints filed by the people through the I4C portal. Due to the increase in online scams, phishing, harassment, and identity theft, it is crucial to quickly categorize and prioritize fraud complaints. The manual processing of these complaints is both time-consuming and difficult, especially as the volume continues to rise. Our solution aims to address this issue by creating a model that can rapidly classify complaints by identifying essential details and patterns, and predict the respective subcategories and categories based on it.

Our model has been designed to handle multiple subcategories for a complaint since the primary category may be further resolved into multiple subcategories. For this, we have explored quite a few approaches, involving but not limited to usage of BERT(Bidirectional Encoder Representations from Transformers), BiLSTM(Bidirectional Encoder Representations from Transformers), etc. and have shared our results for the same. All these approaches were chosen taking into consideration their proven effectiveness in text-classification and complex language tasks.

Once developed, the model can be integrated into a user-friendly dashboard for cybersecurity analysts. This dashboard will visualize complaint types, and sub-types, making tracking trends easy and prioritizing cases needing immediate attention. By automating this classification process, our solution aims to help cybersecurity teams respond more effectively and gain insights into evolving fraud trends. Ultimately, this approach empowers cybersecurity efforts, helping reduce the impact of fraud through faster, more informed decision-making.

### 1.3. Results and Achievements

1. Developed a robust NLP-based model that automatically classifies and analyzes fraud-related complaints, significantly decreasing manual processing time.
2. Achieved an accuracy rate of 53% for the model, indicating good performance in identifying and categorizing various types of fraud.
3. Implemented multi-label classification using advanced Transformer and BiLSTM models, enhancing precision, recall, and F1 score for more effective fraud detection.
4. Improved decision-making by automating the complaint classification process, enabling cybersecurity teams to prioritize cases more effectively.
5. Trained the model on a diverse dataset of over 92,000 complaints, ensuring its capability to manage a wide range of fraud-related scenarios.
6. Delivered insights into emerging fraud trends, supporting resource allocation and strategic planning for cybersecurity teams.

### 1.4. Impact

The development of the NLP-based model for classifying and analyzing fraud-related complaints can be integrated into a broader framework for real-time monitoring, where complaints are instantly categorized and flagged for follow-up based on critical criteria, such as high financial loss. This integration allows for the visualization of complaint types, severity, and affected parties, providing valuable insights for cybersecurity teams. Consequently, teams can address issues more rapidly and allocate resources more efficiently, enhancing their overall response to the growing threat of online fraud. This proactive approach empowers cybersecurity efforts and fosters improved decision-making in combating fraudulent activities.

## Part 2: Technical Approach

---

### 2.1. Data Acquisition and Preprocessing

Data Acquisition: The data acquisition process involved sourcing a structured dataset of cybercrime-related complaints provided in CSV format. Each entry in the dataset included a detailed description of the complaint, captured in a "*crimedatainfo"* column, alongside its corresponding subcategories and overarching categories. These subcategories spanned a wide range of cybercrime activities, including online financial fraud, phishing, cyberbullying, and others. This diversity of complaints ensured the dataset was representative of real-world scenarios, laying the foundation for building a robust machine-learning model.

Data Preprocessing: To prepare the data for machine learning, several preprocessing steps were implemented:

1. Text cleaning:
   a. Normalization: Converted text to lowercase to maintain consistency.
   b. Character Removal: Removed punctuation, special characters, and numbers that were irrelevant to classification.
   c. English and Hingslish Stopword Removal: Eliminated common stopwords like "and" and "the," as they did not contribute meaningful information to the classification task. We found a txt file that contained keywords for both languages and used it to help us  with this task.
   d. Remove short complaints: Complaints with less than 30 characters were removed to prevent noise.
2. Tokenization: Split text into individual tokens (words) to make it processable by the model. This was achieved using NLTK and other similar libraries.
3. Placeholder Substitution for URLs and Email Addresses: Replaced these with placeholder tokens like URL or EMAIL to focus on the context rather than specific identifiers.

4. LOCATION Substitution: LOCATION tags were used to substitute names of places and then to even further refine it, all text between the first and last LOCATION tag(which usually contained the name of more locations) of every complaint was reduced to just 1 tag.
5. Bring categories with no sub-category all under one category: We observed that categories 'Sexually Obscene material', 'Sexually Explicit Act', 'RapeGang Rape RGRSexually Abusive Content', 'Child Pornography CPChild Sexual Abuse Material CSAM' for both datasets lacked any sub-categories so we aggregated all of them under a new category of ‘Sex Abuse’.
6. UPI substitution: We observed the occurrence of a lot of upi apps(like PhonePe, GPay, PayTM, etc) so we aggregated all of them under the special token UPIAPP.
7. Expanding Contractions: The text was split into individual words, and each word was checked against the dictionary. If the word is found in the dictionary, it is replaced with its expanded form; otherwise, it remains unchanged.
8. Singularization: It is the process of splitting the text into individual words and checking each word. If a word ends with an "s" and its singular form (a word without the trailing "s") exists in the vocabulary, the plural form is replaced with the singular. Otherwise, the word remains unchanged.
9. Fixing data leakage:

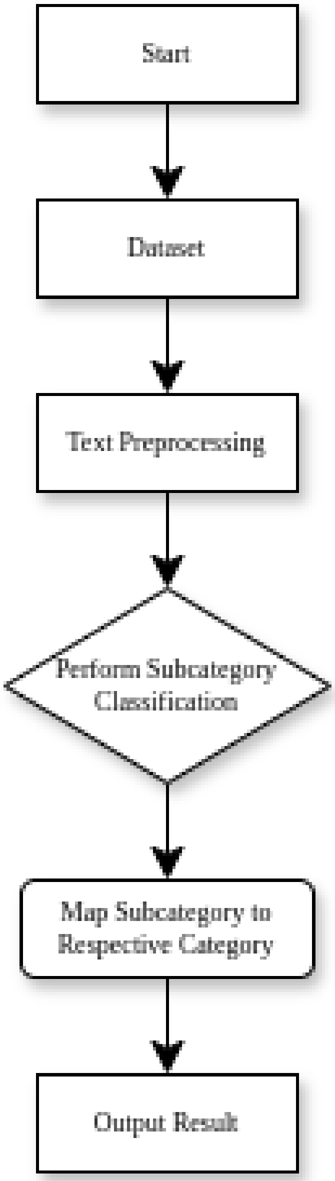
   We fixed 2 errors in the test dataset amongst the others mentioned above:

   a. The train and test datasets had common entries, so we removed the common entries from the train dataset and saved them for testing.
   b. ‘Crime Against Women & Children’ was a category in the test dataset which was a) not in the train dataset, and b) had only 4 entries whose complaints almost matched. So we renamed the category to ‘Hacking Damage to computer system etc’ and the sub-category to ‘Email Hacking’ as it was more appropriate for that.
10. Feature Extraction: Leveraged techniques like Word2Vec(an NLP technique that uses a shallow neural network to learn the meaning of words from a large corpus of text) to convert textual descriptions into numerical representations. These features captured both the frequency and contextual importance of words within the text.

11. Approach-centered data preprocessing: As we mentioned that we made use of multiple approaches like using BERT/RoBERTa (Transformer-based) and BiLSTM, etc. we made use of text preprocessing techniques that enabled effective integration of the above mentioned models for our purpose.

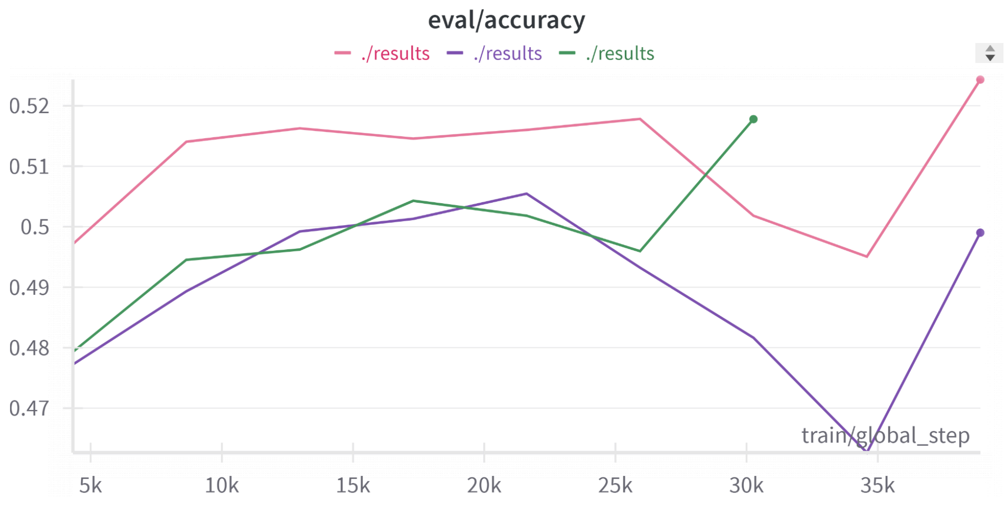
This multi-step preprocessing pipeline ensured that the textual data was clean, meaningful, and appropriately structured, optimizing the input for subsequent machine-learning models.

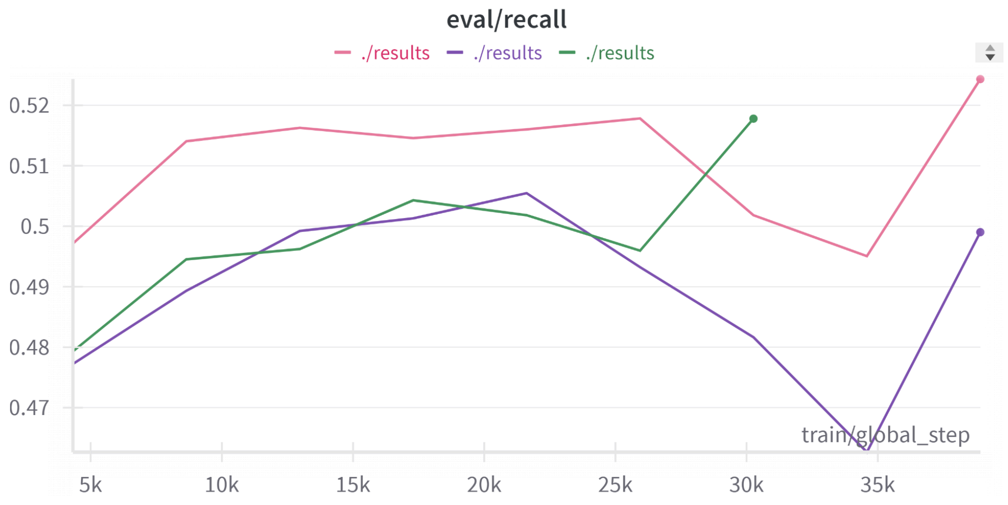
## 2.2. Model Architecture:

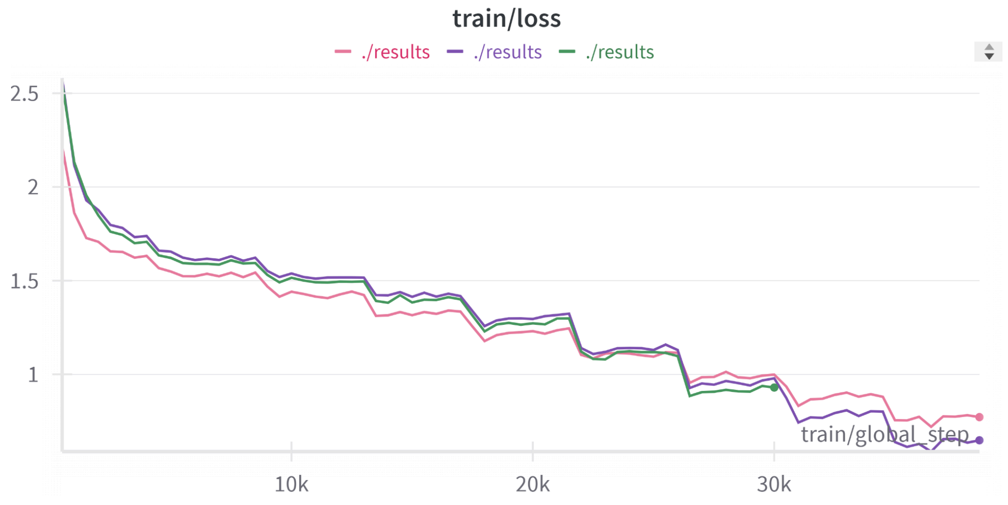

Start
Dataset
Text Preprocessing
Perform Subcategory Classification
Map Subcategory to Respective Category
Output Result

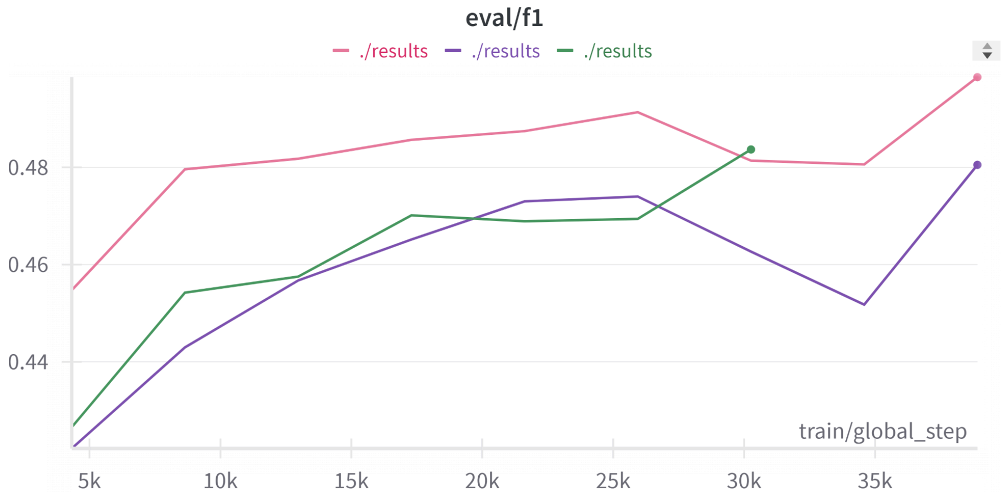
## 2.3. Model Evaluation:

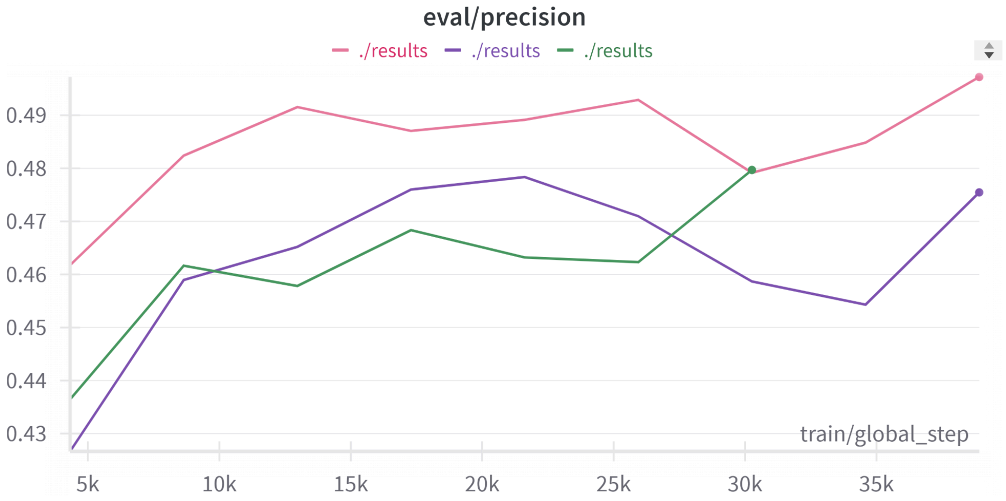
Pink: mpnet-base-v2
Green: ROBERTa
Purple: indicBERT


eval/accuracy


eval/recall


train/loss





## Part 3: Future Work and Potential Improvements

---

### 3.1. Enhancement

Future enhancements for the cybercrime classification model could include incorporating additional contextual features, such as user metadata (e.g., timestamps, geographic location, and device type). These features can provide critical context to improve classification accuracy, particularly for location-specific or device-dependent cybercrime patterns. Moreover, advanced learning techniques like transfer learning (utilizing pre-trained models on similar tasks) and

active learning (iteratively improving the model with targeted samples) can be employed to make the model more adept at handling underrepresented subcategories.
Also, considering that the model accuracy is average for now, we can maybe facilitate better capturing of data and apply advanced preprocessing techniques like transliteration of hinglish text to further improve the classification process.

### 3.2. Scalability

To optimize the solution for large-scale applications, the model can be adapted to handle high-volume data streams with reduced inference latency. This could involve transitioning to more scalable architectures, such as distributed transformer models or on-device inference using lightweight frameworks. Additionally, integrating real-time data processing pipelines would allow the model to classify cybercrime complaints dynamically, enabling real-time response capabilities.

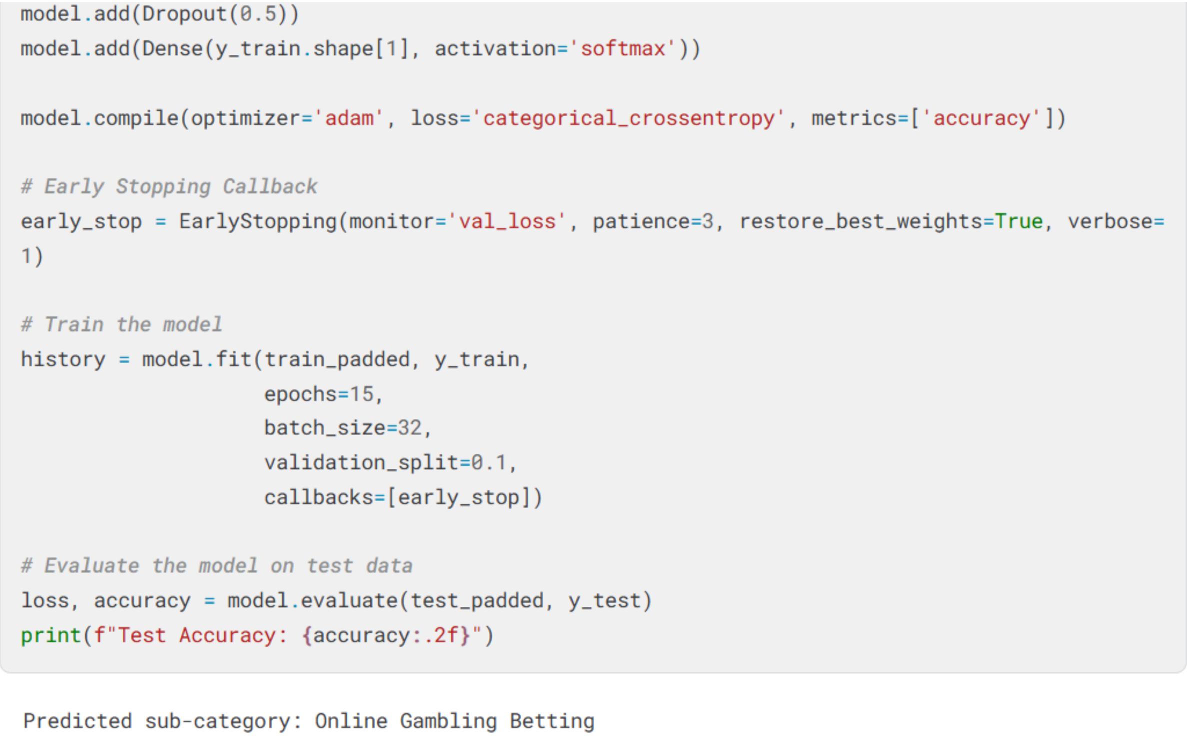
### 3.3. Ethical Concerns

Ethical considerations are paramount when dealing with sensitive data related to cybercrimes:

1. Bias Mitigation: Regular audits will be conducted to identify and address potential biases in the training data or the model's predictions to ensure fairness and equitable treatment across all subcategories.
2. Privacy Protection: Strong data anonymization and encryption measures will be implemented to protect personally identifiable information during data handling and model training.
3. Transparency and Accountability: The model's decision-making process will be designed to be interpretable, ensuring that stakeholders understand the reasons behind classifications. This will foster trust and enable corrections in case of errors.

## Part 4: Results

### 4.1. Output



```
model.add(Dropout(0.5))
model.add(Dense(y_train.shape[1], activation='softmax'))

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Early Stopping Callback
early_stop = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True, verbose=
1)

# Train the model
history = model.fit(train_padded, y_train,
                    epochs=15,
                    batch_size=32,
                    validation_split=0.1,
                    callbacks=[early_stop])

# Evaluate the model on test data
loss, accuracy = model.evaluate(test_padded, y_test)
print(f"Test Accuracy: {accuracy:.2f}")
```

```
Predicted sub-category: Online Gambling Betting
```

4.2. Conclusion: The proposed cybercrime classification model achieved an overall accuracy of 0.53. This demonstrates its ability to process and classify textual descriptions of cybercrime complaints. However, challenges remain in improving recall and precision for underrepresented or ambiguous subcategories, which could benefit from further model refinement.

The solution highlights its potential in cybercrime prevention and mitigation by automating the categorization process and enabling quicker identification of critical threats. Future efforts could focus on integrating transfer learning and active learning to enhance model performance, especially for underrepresented categories. Additionally, exploring more advanced architectures like transformer-based models could unlock greater contextual understanding, while scalability enhancements will ensure the model's readiness for real-time deployment across large datasets.

Ethical considerations, including the mitigation of biases in data and predictions and the protection of sensitive information, will remain central to the model's development. Ensuring fairness, transparency, and reliability will be critical as this solution evolves to address real-world cybercrime challenges. With these improvements, the model has the potential to significantly enhance the speed and accuracy of cybercrime analysis, aiding law enforcement and cybersecurity professionals in their efforts to combat emerging threats.