Project Report

Submission for IndiaAI CyberGuard AI Hackathon

Team Name: Team Garuda

Parth Thakre, Shreeharsh Shivpure, Atharva Bhajan, Aditi Patil, Aishwary Gathe

Date of Submission: 06-Nov-2024

Content

Chapter 1

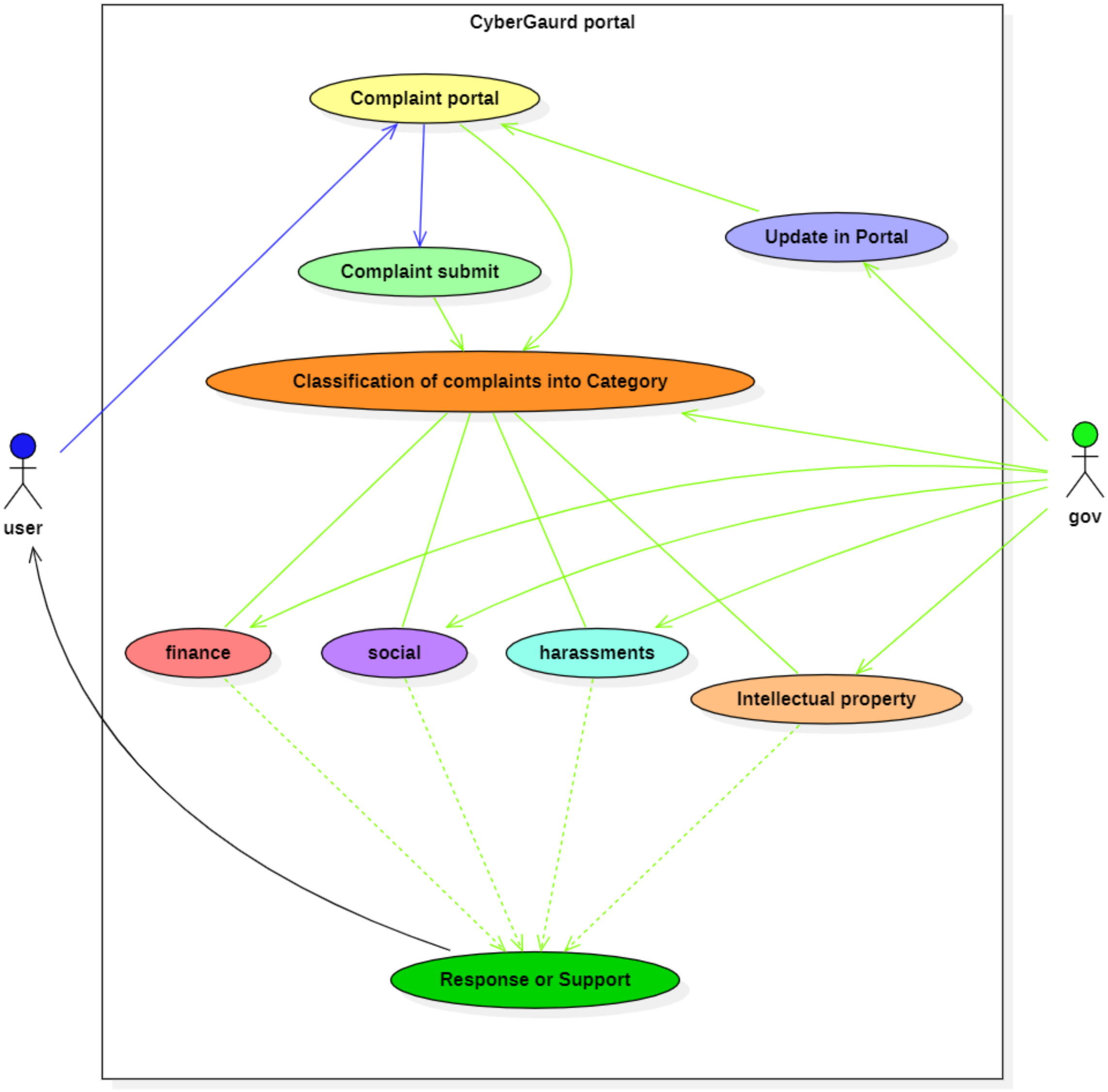
1.1 Problem Statement: Development of an NLP Model for Text Analytics and Classification.

1.2 Solution Overview:

For the IndiaAI Cyberguard AI Hackathon, our project centers around building an NLP-based model that automatically classifies and analyzes fraud-related complaints. Given the rise in online scams, phishing, and identity theft, there is a pressing need to categorize and prioritize fraud complaints quickly. Manually processing these complaints is time-consuming and challenging, especially as the volume grows. Our solution seeks to change that by developing a model that can instantly classify complaints by identifying key details like the type of fraud, the type of victim, and other relevant factors, such as the severity and urgency of each case.

Our classification model will be designed to handle multiple categories within a single complaint since some cases involve more than one type of fraud. For this, we'll explore advanced multi-label classification models, particularly using Transformers, as they're excellent for text analysis and have proven effective in complex language tasks. We have trained and fine-tuned the model to ensure accuracy, optimizing for key metrics like precision, recall, and F1 score.

Once developed, the model can be integrated into a user-friendly dashboard for cybersecurity analysts. This dashboard will visualize complaint types, victim categories, and severity, 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 fraud's impact through faster, more informed decision-making.\



1.2.1 Use Case Diagram

1.3 Key Results and Achievements

1. Developed a robust NLP-based model that automatically classifies and analyzes fraud-related complaints, reducing manual processing time.
2. Achieved a model accuracy of 86%, demonstrating high performance in identifying and categorizing various fraud types.
3. Implemented multi-label classification using advanced Transformer models, optimizing precision, recall, and F1 score for better fraud detection.

4. Improved decision-making by automating the complaint classification process, allowing cybersecurity teams to prioritize cases efficiently.
5. Model trained on a diverse dataset of over 92,000 complaints, ensuring it can handle fraud-related cases.
6. Achieved strong performance in detecting online financial fraud with a precision of 0.86 and recall of 1.00, enabling faster response times.
7. Provided insights into emerging fraud trends, aiding resource allocation and strategic planning for cybersecurity teams.

Accuracy	0.86
Number of Complaints	92,000
Categories of Classification	14
Fraud Categories	9

1.4 Impact

Think about integrating the model within a broader framework for real-time monitoring, where complaints are classified instantly and flagged for follow-up if they meet critical criteria (e.g., high financial loss). Visualizing complaint types, severity, and affected parties can offer valuable insights for cybersecurity teams, enabling them to address issues faster and allocate resources more effectively.

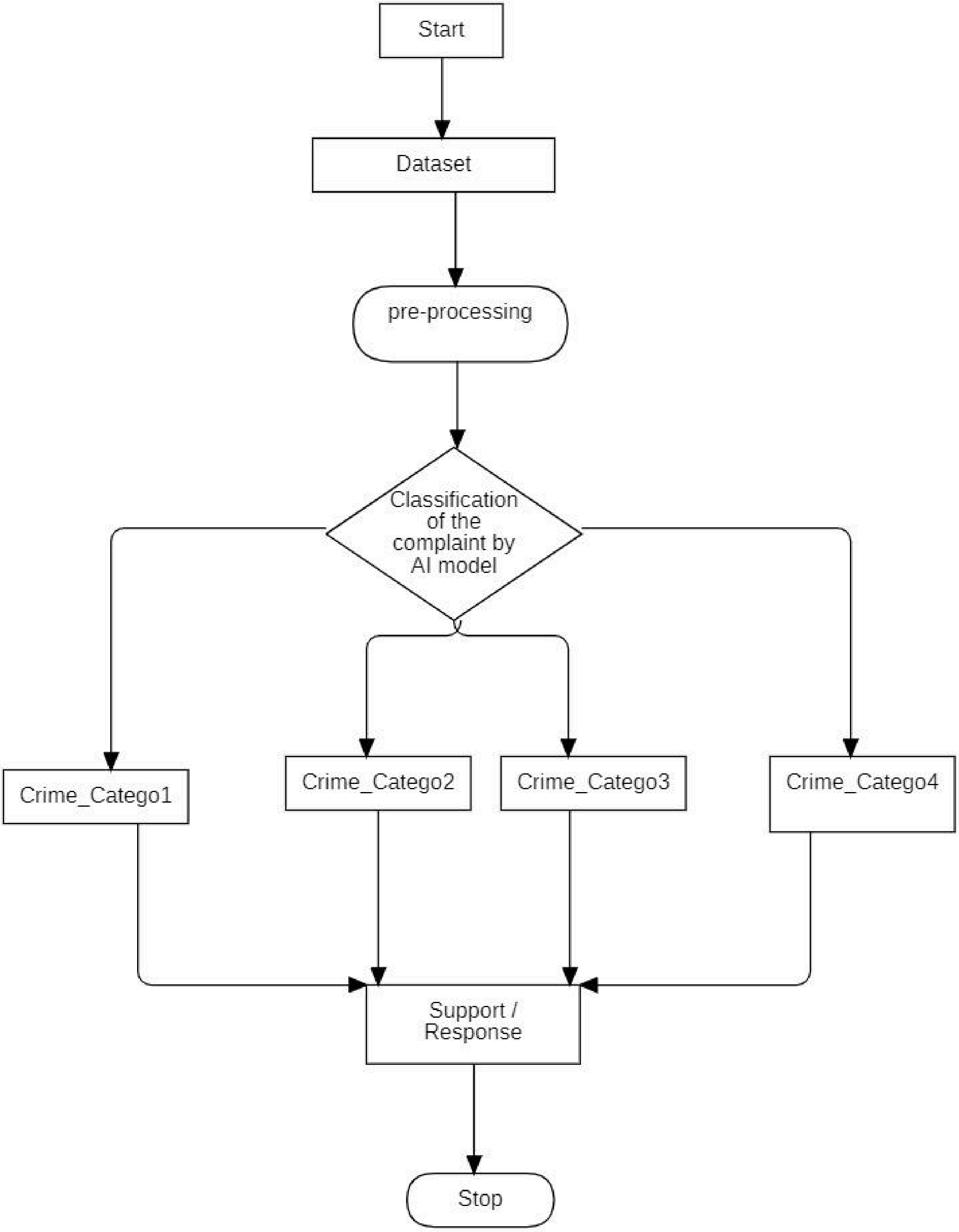
Chapter 2

2.1 Data Acquisition and Preprocessing

Data Acquisition: The data acquisition process involved obtaining a comprehensive Excel sheet, which already contained a structured dataset of cybercrime-related complaints. This dataset was categorized into various types of cybercrimes, such as online financial fraud, cyber attacks, child pornography, cryptocurrency crimes, and others. Each entry in the sheet was associated with specific details related to the type of complaint, including relevant descriptions or indicators. The categories represented a wide range of criminal activities, providing a diverse set of data points for analysis. Acquiring this data was the initial and crucial step, as it provided a well-organized base for the next stages of preprocessing and model development. The dataset was designed to be cleaned, tokenized, and processed for building a machine-learning model capable of classifying and identifying different types of cybercrimes accurately.

Data Preprocessing: Data preprocessing is essential to ensure that a machine-learning model receives clean and When preparing data for a machine-learning model, it's important to detail the preprocessing steps taken to ensure clean and meaningful input. For text data, we first perform text cleaning. This involves converting all text to lowercase to ensure consistency, removing punctuation and special characters that are irrelevant, and eliminating stopwords like "the" and "is," which do not contribute to fraud detection. We also apply lemmatization to reduce words to their base forms, ensuring they retain contextual meaning (for example, turning "buying" into "buy" and "better" into "good"). Any URLs, email addresses, or numbers in the text that don't contribute to fraud classification are replaced with placeholder tokens like <URL>. After cleaning, we tokenize the text, splitting it into individual words or phrases using functions such as NLTK's word_tokenize(). These steps are crucial for ensuring the model receives input that is both clean and structured for optimal performance.

2.2 Model Architecture



2.3 Model Evaluation

Categories	precision	recall	f1-score	support
any other cybercrime	0.98	0.68	0.80	10727
cryptocurrency crime	0.99	0.70	0.82	477
cyber attack/ dependent crimes	0.98	1.00	1.00	3608
cyber terrorism	0.94	0.64	0.76	160
hacking damage to computer-computer system etc	0.99	0.68	0.80	1682
online and social media-related crime	0.75	0.85	0.79	11877
online cyber trafficking	0.97	0.48	0.64	180
online financial fraud	0.86	1.00	0.92	56718
online gambling betting	0.98	0.56	0.72	438

Chapter 3

3.1 Enhancement

Future improvements can focus on incorporating additional features such as user metadata (e.g., location, device information) to help classify cybercrimes more effectively. Refining the training process using techniques like active learning and transfer learning could help the model learn better from limited data, particularly in underrepresented categories.

3.2 Scalability

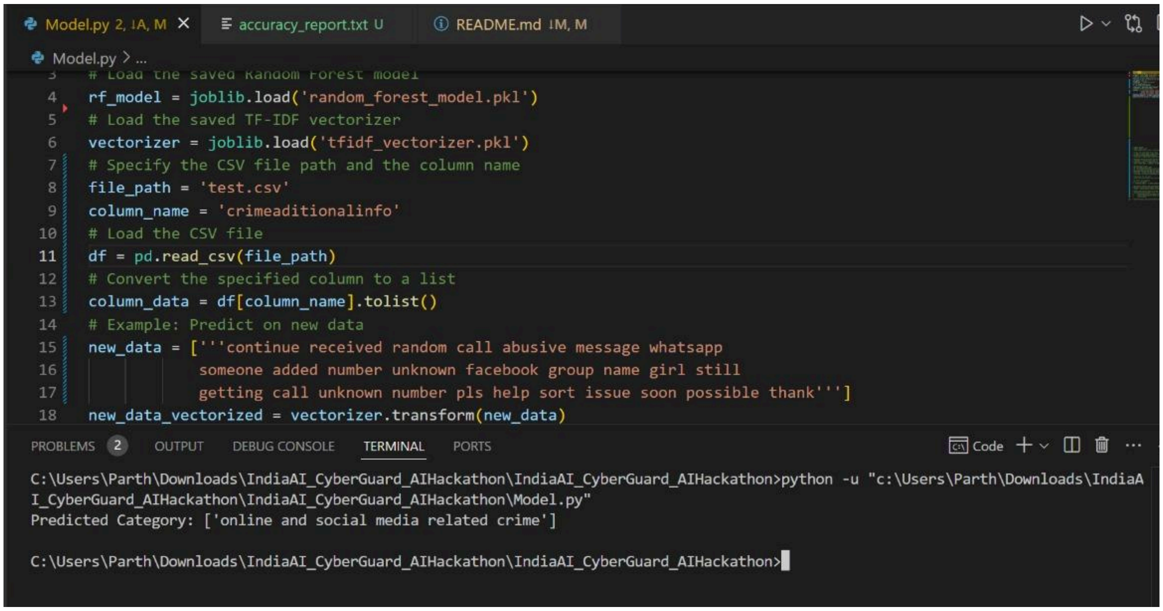
The solution can be optimized for large-scale deployments by improving its efficiency in terms of inference time and resource usage. This could involve using more scalable architectures, distributing processing workloads, and enhancing model performance for real-time processing in vast datasets.

3.3 Ethical Considerations

There are ethical concerns regarding data bias and privacy when classifying sensitive topics like cybercrime. Efforts will be made to ensure fairness by auditing the model regularly for bias and ensuring that the training data does not disproportionately represent or neglect certain categories. Additionally, privacy measures must be implemented to safeguard personal information while processing the data.

Chapter 4

4.1 Output



```python
# Load the saved Random Forest model
rf_model = joblib.load('random_forest_model.pkl')
# Load the saved TF-IDF vectorizer
vectorizer = joblib.load('tfidf_vectorizer.pkl')
# Specify the CSV file path and the column name
file_path = 'test.csv'
column_name = 'crimeaditionalinfo'
# Load the CSV file
df = pd.read_csv(file_path)
# Convert the specified column to a list
column_data = df[column_name].tolist()
# Example: Predict on new data
new_data = ['''continue received random call abusive message whatsapp
            someone added number unknown facebook group name girl still
            getting call unknown number pls help sort issue soon possible thank''']
new_data_vectorized = vectorizer.transform(new_data)
```

```
C:\Users\Parth\Downloads\IndiaAI_CyberGuard_AIHackathon\IndiaAI_CyberGuard_AIHackathon>python -u "c:\Users\Parth\Downloads\IndiaA
I_CyberGuard_AIHackathon\IndiaAI_CyberGuard_AIHackathon\Model.py"
Predicted Category: ['online and social media related crime']

C:\Users\Parth\Downloads\IndiaAI_CyberGuard_AIHackathon\IndiaAI_CyberGuard_AIHackathon>
```

4.2 Conclusion

The model has demonstrated an overall accuracy of 0.86, performing well in categories like "online financial fraud" and "cyber attacks." However, there is room for improvement, especially in categories with low recall and precision. The solution has significant potential for cybercrime prevention and mitigation by automating classification and providing insights for quicker threat detection. Future work could focus on enhancing the model through advanced techniques like active learning, exploring more complex models, and addressing scalability for large-scale deployments. Ethical considerations, including bias and privacy, will also be crucial as the model evolves to ensure fairness and reliability in real-world applications.